U. S. ARMY RESEARCH OFFICE

Report No. 84-1

February 1984

TRANSACTIONS OF THE FIRST ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Host

U. S. Army research Office

held at

George Washington University

Washington, D. C.

Approved for public release; distribution unlimited. The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

FOREWORD

The first, in the series of annual meetings entitled Army Conferences on Applied Mathematics and Computing, was held on 9-11 May 1983 at George Washington University, Washington, DC. These meetings, sponsored by the Army Mathematics Steering Committee (AMSC), combines two symposia, namely the Conferences of Army Mathematicians and the Numerical Analysis and Computers conferences. The joining of these two meetings underscores the growing synergism between applicable mathematics and computing. In the next two paragraphs short histories of the two forerunners of these new conferences are presented.

The Office of Ordnance Research, now the Army Research Office, planned a series of symposia called Ordnance Conferences of Arsenal Mathematicians. The first of these meetings was held on 29 October 1954 at Watertown Arsenal. The purpose of these meetings was to focus the attention of applied mathematicians on some specific ordnance projects needing high level scientific knowledge. Following the Seventh Conference of Ordnance Mathematicians, the AMSC became the sponsor for these symposia. This committee requested that these conferences be held on an Army-wide basis, and that their name be changed to Conferences of Army Mathematicians. This name continued through the twenty-eight conference, which was held on 28-30 June 1982 at the Uniformed Services University of Health Sciences, Bethesda, Maryland.

About five years after the first Conference of Arsenal Mathematicians was held, the Office of Ordnance Research organized an OOR Liason Group on Computers. Two of these meetings were held, one in 1959 and the other in 1960, to exchange information of interest to managers of ordnance computers. [These gatherings can be considered as the forerunners of the Numerical Analysis and Computers Conferences.] The AMSC decided that these meetings should be conducted on an Army-wide basis and be entitled the ARO Working Groups on Computers. Their purpose was to provide a format for exchanging ideas on the Army's desires, capabilities, and interest in the field of 'other-than business' application of computers, and they should provide AMSC and ARO with information on the Army's need for computers, requirements for assistance in research in numerical analysis, and other kinds of mathematics. Two meetings, one in 1962 and the other in 1964, were held under the above mentioned title. Starting in 1965 these symposia were held, except for 1973, on a yearly basis, and at first were entitled Army Numerical Analysis Conferences. Starting with the 1975 meeting they became known as the Numerical Analysis and Computers Conferences. The last meeting in this series was held on 3-4 February 1982 at the U. S. Army Engineer Waterways Experiment Station, Vicksburg, Mississippi.

Members of the AMSC would like to thank Professor Nozer D. Singpurwalla of George Washington University, Washington, DC, for his invitation to hold the first meeting of this new series of conferences at his university. They would also like to thank him for providing such excellent facilities for the symposium, and for all the help he and other members of this staff provided to insure a smooth running meeting.

The Program Committee was especially pleased with the quality of the contributed papers for this meeting. Also the number of these papers, 57, helped get these new conferences off to a good start. There was a Special Session on Distributed Command and Control. The speakers in this session covered topics of special interest to various Army groups. A list of the invited speakers along with the titles of their addresses is noted below. Members of the Committee are sorry to report that, due to a previous engagement, Dr. Karl J. Astrom of the Lund Institute of Control, Sweden, was unable to accept our invitation to address this conference.

Speaker and Affiliations	Title
Professor Julian Cole Rensselaer Polytechnic Institute	Pertubation Techniques for Fluid Dynamics
Professor Andrew Majda University of California-Berkeley	Stability and Instability for Shock Waves
Professor Sanjoy K. Mitter Massachusetts Institute of Technology	Adaptive Controls
Dr. M. Yousuff Hussaini NASA Langley Research Center	Spectral Methods for Partial Differential Equations
Professor Carl de Boor Mathematics Research Center, University of Wisconsin-Madison	Multivariate B-Splines

Members of the AMSC would like to thank the speakers and all the other individuals who contributed to the success of this meeting. They have requested that most of the contributed papers be made available in printed form. These research articles will enable many persons that could not attend the smposium to profit by these contributions to the scientific literature.

TABLE OF CONTENTS*

*This Table of Contents lists only the papers that are published in this Technical Manual. For a list of all the papers presented at the First Army Conference on Applied Mathematics and Computing, see the Adjenda.

FIRST ARMY CONFERENCE

ON

APPLIED MATHEMATICS AND COMPUTING

9-11 May 1983

GEORGE WASHINGTON UNIVERSITY
WASHINGTON, D.C.
2201 G STREET
BUILDING C

******************** AGENDA ********************

Monday, 9 May 1983

0815-0845 REGISTRATION - Building C, Lobby

0845-0900 OPENING REMARKS

0900-1000 GENERAL SESSION I - Auditorium, C-108
CHAIRMAN - Billy Z. Jenkins, US Army Missile Command
PERTURBATION TECHNIQUES FOR FLUID DYNAMICS
Julian Cole, Rensselaer Polytechnic Institute

1000-1020 BREAK

Monday, 9 May 1983

1020-1200 TECHNICAL SESSION I - C-222

CHAIRMAN - Joseph Matta, Chemical Systems Laboratory, ARRADCOM

SMALL DISTURBANCE THEORY FOR THE TRANSONIC FLOW OF DUSTY GAS

Donald A. Drew, Rensselaer Polytechnic Institute and Fredrick J. Zeigler, General Dynamics

INSTABILITY DUE TO VISCOUS STRATIFICATION IN PIPE FLOW

D. D. Joseph, University of Minnesota, and M. Renardy and Y. Renardy, Mathematics Research Center, University of Wisconsin-Madison

NUMERICAL SIMULATION OF THE FLOW OF A CHEMICAL CONTAMINANT SUBJECTED TO A JET IMPINGEMENT

L. M. Chang, US Army Ballistic Research Laboratory, ARRADCOM

UNIQUENESS AND STABILITY OF FLOW OF IMMISCIBLE FLUIDS OF DIFFERENT VISCOSITY

D. D. Joseph, K. Nguyen, and G. Beavers, University of Minnesota

FLOW PAST A FLEXIBLE MEMBRANE

Jean-Marc Vanden-Broeck, Mathematics Research Center, University of Wisconsin-Madison

1200-1330 LUNCH

Monday, 9 May 1983

1020-1200 TECHNICAL SESSION II - C-220

CHAIRMAN - James Thompson, US Army Tank-Automotive Command

TRANSIENT VEHICLE DYNAMICS DUE TO FLEXIBLE CHASSIS-SUSPENSION INTERACTION

E. J. Haug, A. A. Shabana, and S. S. Kim, University of Iowa

MACSYMA AND MECHANICAL SYSTEMS

M. A. Hussain, General Electric Co. and Ben Noble, Mathematics Research Center, University of Wisconsin-Madison

ON THE CONVERSION OF LIN PACK TO ADA

Benjamin J. Martin, Atlanta University and Robert Bozeman, Morehouse College

AUTOMATIC GENERATION OF TAYLOR SERIES IN PASCAL-SC: BASIC OPERATIONS AND APPLICATIONS TO ORDINARY DIFFERENTIAL EQUATIONS

L. B. Rall, Mathematics Research Center, University of Wisconsin-Madison and George Corliss, Marquette University

HAND HELD PROGRAMMABLE CALCULATORS - THE CHALLENGE OF A KEPLERIAN

Donald I. Thompson, US Army, White Sands Missile Range

1200-1330 LUNCH

Monday, 9 May 1983

1330-1530 TECHNICAL SESSION III - C-222

CHAIRMAN - Clarence W. Kitchens, Jr., US Army
Ballistic Research Laboratory, ARRADCOM

STRUCTURAL EVALUATION OF A DISCONTINUOUS SHELL SUBJECTED TO AN INTERNAL BLAST

Aaron Das Gupta and Henry L. Wisniewski, US Army Ballistic Research Laboratory, ARRADCOM

COMPUTATION OF SHOCKS IN 2-D GAS DYNAMICS

James Glimm, New York University

RIEMANN SOLVERS, THE ENTROPY CONDITION AND HIGH RESOLUTION DIFFERENCE APPROXIMATIONS

Stanley Osher, University of California

ARTIFICIAL MASS CONCEPT AND TRANSONIC VISCOUS FLOW EQUATION

George S. Dulikravich, University of Texas-Austin

V-STATES. STEADY-STATE SOLUTIONS OF THE EULER EQUATIONS, THEIR STABILITY & NONLINEAR EVOLUTION

Norman J. Zabusky and Edward A. Overman, II, University of Pittsburgh

DIAGNOSTIC ALGORITHMS FOR CONTOUR DYNAMICS

Edward A. Overman, II and Norman J. Zabusky, University of Pittsburgh

1530-1550 BREAK

1550-1650 GENERAL SESSION II - C-108

CHAIRMAN - Raymond Sedney, US Army Ballistic
Research Laboratory, ARRADCOM

STABILITY & INSTABILITY FOR SHOCK WAVES

Andrew Majda, University of California-Berkeley

Monday, 9 May 1983

1330-1530 TECHNICAL SESSION IV - C-220

CHAIRMAN - Norman P. Coleman, Jr., US Army
Armament Research & Development Command

GROUP ANALYSIS OF THE VON KÁRMÁN EQUATIONS

W. F. Ames, Georgia Institute of Technology and K. A. Ames, Iowa State University

EXISTENCE OF QUASI-SOLUTIONS OF NONLINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS

G. S. Ladde, V. Lakshmikantham, University of Texas-Arlington, and A. S. Vatsala, Bishop College

ON THE CONTROL OF A LINEAR STOCHASTIC SYSTEM WITH FINITE HORIZON

P. L. Chow and J. L. Menaldi, Wayne State University

A NONLINEAR INTEGRAL EQUATION OCCURRING IN A SINGULAR FREE BOUNDARY PROBLEM

Klaus Hollig and John A. Nohel, Mathematics Research Center, University of Wisconsin-Madison

NONLINEAR INVERSE HEAT TRANSFER CALCULATIONS IN GUN BARRELS

Alfred S. Carasso, National Bureau of Standards

ON THE EXTREMUM OF BILINEAR FUNCTIONAL FOR HYPERBOLIC TYPE P.D.E.

C. N. Shen, Benet Weapons Laboratory, ARRADCOM

1530-1550 BREAK

1550-1650 GENERAL SESSION II - C-108

CHAIRMAN - Raymond Sedney, US Army Ballistic
Research Laboratory, ARRADCOM

STABILITY & INSTABILITY FOR SHOCK WAVES

Andrew Majda, University of California-Berkeley

Tuesday, 10 May 1983

0830-1010 TECHNICAL SESSION V - C-222

CHAIRMAN - C. N. Shen, Benet Weapons Laboratory, ARRADCOM

NONLINEAR MULTIDIMENSIONAL MODELS, LEAST SQUARES AND REDUCED RESIDUALS

Aivars K. Celmins, US Army Ballistic Research Laboratory, ARRADCOM

APPLICATION OF THE PRINCIPAL COMPONENT METHOD TO TRAJECTORY ESTIMATION

William S. Agee and Robert H. Turner, US Army White Sands Missile Range

OPTIMAL REDUCED ORDER CONTROLLERS FOR DISCRETE-TIME LINEAR SYSTEMS WITH PARTIAL STATE ESTIMATION

Maurice F. Hutton, US Army White Sands Missile Range

INFRARED MODELLING OF HIGH VALUE FIXED TARGETS

Richard A. Weiss, Waterways Experiment Station, Corps of Engineers

ORTHOGONAL SCHEMES FOR STRUCTURAL OPTIMIZATION

R. J. Plemmons, M. W. Berry, North Carolina State University and Robert C. Ward and M. T. Heath Union Carbide Corporation

1010-1030 BREAK

Tuesday, 10 May 1983

0830-1010 TECHNICAL SESSION VI - C-220

CHAIRMAN - Dennis M. Tracey, US Army Materials &
Mechanics Research Center

THREE DIMENSIONAL STRESS ANALYSIS OF A PINNED-JOINT IN A PROJECTILE

Tien-Yu Tsui, US Army Materials & Mechanics Research Center, and M.L. Chiesa and M. L. Callabresi, Sandia National Laboratories

A SIMPLE APPROACH FOR DETERMINATION OF BURSTING PRESSURE OF A THICK-WALLED CYLINDER

S. C. Chu, US Army Armament Research and Development Command

STRAIN-HARDENING EFFECT ON STRESS-INTENSITY FACTORS FOR RADIAL CRACKS IN AN AUTOFRETTAGED GUN BARREL

S. L. Pu and P.C.T. Chen, Benet Weapons Laboratory, ARRADCOM

STRESS DISTRIBUTION IN A CYLINDRICAL BAR SUBJECTED TO CYCLIC TORSIONAL LOADING

P.C.T. Chen, Benet Weapons Laboratory, ARRADCOM, and H.C. Wu, University of Iowa

FINITE ELEMENT RESULTS OF PRESSURIZED THICK TUBES BASED ON TWO ELASTIC-PLASTIC MATERIAL MODELS

P.C.T. Chen and G.P. O'Hara, Benet Weapons Laboratory, ARRADCOM

1010-1030 BREAK

Tuesday, 10 May 1983

1030-1210 TECHNICAL SESSION VII - C-222

CHAIRMAN - James A. Schmitt, US Army, Ballistic
Research Laboratory, ARRADCOM

ON THE SOLUTION OF ELLIPTIC FINITE ELEMENT EQUATION

Seymour V. Parter, University of Wisconsin-Madison

A FINITE INTEGRAL TRANSFORM (FIT) METHOD FOR DIFFERENTIAL EQUATIONS HAVING ASYMPTOTIC SOLUTIONS

Charles J. Daly, Cold Regions Research and Engineering Laboratory

A FINITE DIFFERENCE METHOD FOR ANY PARTIAL DIFFERENTIAL EQUATION

R. Yalamanchili, US Army Armament Research and Development Command

NUMERICAL BOUNDARY CONDITIONS FOR FLOW PROBLEMS

M. D. Gunzburger and G. J. Fix, Carnegie-Mellon University

LEAST SQUARES APPROXIMATIONS TO COMPRESSIBLE FLOW PROBLEMS

G. J. Fix and M. D. Gunzburger, Carnegie-Mellon University

1210-1340 LUNCH

Tuesday, 10 May 1983

1030-1210 TECHNICAL SESSION VIII - C-220

CHAIRMAN - Bert Zarwyn, US Army Electronics
R & D Command

GEOMETRIC PROGRAMING

Patrick D. Allen, US Army Concepts Analysis Agency and David W. Baker, Ghetty Oil Company

SPACE AND TIME ANALYSIS IN DYNAMIC PROGRAMING ALGORITHMS

Bennett Setzer, Atlanta University

MATHEMATICAL ANALYSIS OF THE COUNTERFIRE DUEL: TANKS VS. ANTI-TANK MUNITIONS

Joseph V. Michalowicz, Harry Diamond Laboratories

ALGORITHM FOR CALCULATING UNIT SEPARATION DISTANCES

Timothy M. Geipe, Harry Diamond Laboratories

ON A GENERALIZED RAYLEIGH-RITZ METHOD FOR STRUCTURAL DYNAMICS PROBLEMS IN CONJUNCTION WITH FINITE ELEMENTS

Julian J. Wu, Benet Weapons Laboratory, ARRADCOM

A NUMERICAL TECHNIQUE IN THE HYDRODYNAMIC THEORY OF FOIL BEARINGS

I. G. Tadjbakhsh, Rensselaer Polytechnic Institute, G. Ahmadi, Clarkson College, and E. A. Saibel, US Army Research Office

1210-1340 LUNCH

Tuesday, 10 May 1983

1340-1540 SPECIAL SESSION ON DISTRIBUTED COMMAND AND CONTROL
C-108

CHAIRMAN - Robert L. Launer, US Army Research Office

INTERACTIVE COMBAT SIMULATION AS AN EXPERIMENTAL
TOOL FOR THE EVOLUTION OF C^3I SYSTEMS

*Reiner K. Huber, Hochschule Der Bundeswehr
Federal Republic of Germany*

DECISION ALGORITHMS IN FUZZY SITUATIONS

*Hans-Juergen Zimmerman, Operations
Research, Federal Republic of Germany*

COMMAND AND CONTROL SIMULATION PROBLEMS AND
APPROACHES

Charles Todd, Jet Propulsion Laboratory

ARTILLERY CONTROL ENVIRONMENT: AN EXPERIMENTAL
TOOL

*Jill Smith and Jock Grynovicki, US Army Ballistic
Research Laboratory, ARRADCOM*

1540-1600 BREAK

1600-1700 GENERAL SESSION III - C-108

CHAIRMAN - Stephen Wolff, US Army Ballistic
Research Laboratory, ARRADCOM

ADAPTIVE CONTROLS (Tentative)

*Karl J. Astrom, Lund Institute of Control,
SWEDEN*

Tuesday, 10 May 1983

1340-1540 POSTER SESSION - C-208

MOVING FINITE ELEMENT RESEARCH FOR SHOCK HYDRODYNAMICS, CONTINUUM MECHANICS AND COMBUSTION

Robert J. Gelinas, Said K. Doss, Neil N. Carlson, Science Applications, Inc.

ELASTIC WAVE SCATTERING FROM CYLINDRICAL CAVITIES AND SOLID INCLUSIONS ANALYZED BY THE RESONANCE METHOD

P. P. Delsanto, J.D. Alemar, and E. Rosario, University of Puerto Rico, and A. Nagl and H. Uberall, Catholic University of America

SOLUTION OF A TYPE OF ILL-CONDITIONED SIMULTANEOUS LINEAR EQUATIONS

Shunsuke Takagi, US Army Cold Regions Research and Engineering Laboratory

FINITE ELEMENT ANALYSIS OF FABRICS WITH NONLINEAR BI-AXIAL STRESS-STRAIN LAWS

Arthur R. Johnson, US Army Natick Research and Development Laboratories

1540-1600 BREAK

1600-1700 GENERAL SESSION III - C-108

CHAIRMAN - Stephen Wolff, US Army Ballistic Research Laboratory. ARRADCOM

ADAPTIVE CONTROLS (Tentative)

Karl J. Astrom, Lund Institute of Control, SWEDEN

Wednesday, 11 May 1983

0830-1030 TECHNICAL SESSION IX - C-222

CHAIRMAN - Douglas E. Kooker, US Army Ballistic Research Laboratory, ARRADCOM

STABILITY OF PLANE NEFS (NEAR-EQUIDIFFUSIONAL FLAMES) WITHOUT INVOKING THE CONSTANT-DENSITY APPROXIMATION

T. Jackson and A. Kapila, Rensselaer Polytechnic Institute

FLAMES IN FLUIDS: THEIR INTERACTION AND STABILITY

M. Matalon and B.J. Matkowsky, Technological Institute, Northwestern University

A CALCULATION OF WRINKLED FLAMES

H.V. McConnaughey, Mathematics Research Center, University of Wisconsin-Madison, G.S.S. Ludford, Cornell University, and G.I. Sivashinsky, Tel-Aviv University

EVOLUTION OF NEAR CHAPMAN-JOUGET DETONATIONS

D. S. Stewart, University of Illinois and G.S.S. Ludford, Cornell University

THE DETONATION WAVE RESULTING FROM SHOCK-INDUCED TRANSITION OF A DEFLAGRATION

A. A. Oyediran and G.S.S. Ludford, Cornell University

THERMAL AND TRANSFORMATION STRESSES IN HOLLOW TUBES DURING THE QUENCHING PROCESS

J. D. Vasilakis, Benet Weapons Laboratory, ARRADCOM

1030-1050 BREAK

Wednesday, 11 May 1983

0830-1030 TECHNICAL SESSION X - C-220

CHAIRMAN - James T. Wong, US Army Aviation R & D
Command, Research and Technology
Laboratories

FACTORIZATIONS OF DIAGONALLY DOMINANT OPERATORS ON ℓ_1.

J. D. Ward, Texas A & M University

EXPLICIT FORMULAS FOR C^n PIECEWISE HERMITE INTERPOLANTS

R. W. Soanes, Benet Weapons Laboratory, ARRADCOM

BIVARIATE QUADRATIC SPLINES ON CRISSCROSS TRIANGULATIONS

Charles K. Chui, Texas A & M University

DESIGNING FINITE ELEMENT SOFTWARE FOR LARGE DEFORMATION ANALYSIS

Dennis M. Tracey and Roshdy S. Barsoum, Army Materials and Mechanics Research Center

HIGHLY PARALLEL ARCHITECTURES FOR SOLVING ORDINARY DIFFERENTIAL EQUATIONS

Richard H. Travassos, Integrated Systems Inc.

1030-1050 BREAK

Wednesday, 11 May 1983

1050-1250 GENERAL SESSION IV - C-108

CHAIRMAN - J. Chandra, US Army Research Office

SPECTRAL METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

M. Yousuff Hussaini, Institute for Computer Applications in Science & Engineering

MULTIVARIATE B-SPLINES

Carl de Boor, Mathematics Research Center, University of Wisconsin-Madison

1300- ADJOURN

ASYMPTOTICS AND NUMERICS

Julian D. Cole
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

ABSTRACT. Limit process expansions applied to problems for partial differential equations are discussed. Three different types of non-uniformities, non-linear far field, non-linear local singularity, and singular boundary conditions are illustrated by example. The use of different limits to construct overlapping expansions is emphasized. Matching is sketched briefly.

I. INTRODUCTION. The combination of asymptotic and numerical methods provides a very useful approach to a wide variety of problems involving partial differential equations. In many cases an "exact" problem can be formulated which is non-linear. This non-linear problem in its dimensionless form contains various parameters. When a parameter is effectively small (or large) the dependence of the solution on the parameter can be studied. The dependence is asymptotic and can very often be connected with a limit process in terms of rescaled coordinates and similarity parameters. Very often, too, different expansions are needed to achieve validity in different regions. Expansions in adjacent regions of physical and parameter space can be connected by asymptotic matching. The purpose of the asymptotic analysis is to define the simplest problems illustrating the phenomena. Both qualitative and quantitative understanding is sought. The limit processes used to construct the expansions bring out the similarity parameters of the problem and provide scaling laws.

In problems that are intrinsically non-linear the representative simplified problems retain that feature. Thus numerical work comes into the picture. The asymptotic analysis isolates the crucial numerical problem and can reduce the amount of computations required by an order of magnitude. Typical problems have to be solved numerically in infinite or semi-infinite domains. But asymptotic matching provides the necessary far field boundary conditions and determines the constants. Because of the asymptotic analysis the numerical problems contain the minimum number of parameters. Further the variables and parameters are scaled to O(1), simplifying the numerics.

Limit process expansions are applicable to all problems where the limits of the solutions exist. If, for example, oscillatory functions appear which do not have limits, then multi-scale, WKB, or averaging methods can be adopted. Limit process expansions are considered here. An illustrative form for a function of one variable x, and two parameters ε, μ is the following

$$f(x;\varepsilon,\mu) = \alpha_1(\varepsilon)f_1(\overline{x};\overline{\mu}) + \alpha_2(\varepsilon)f_2(\overline{x};\overline{\mu}) + \ldots \tag{1}$$

for $a \leq x \leq b$. ε is the basic parameter; $\varepsilon \ll 1$. The $\alpha_j(\varepsilon)$ form an asymptotic sequence so that as $\varepsilon \to 0$, $\frac{\alpha_{j+1}}{\alpha_j} \to 0$. $\overline{x} = \overline{x}(x;\varepsilon)$, $\overline{\mu} = \overline{\mu}(\varepsilon,\mu)$ are definite functions chosen for the purposes of the expansion. $\overline{\mu}(\varepsilon,\mu)$ is a similarity parameter. The choices are not arbitrary but are connected with the requirement that the limit process associated with (1) be distinguished (cf [1]). The successive terms in (1) are computed from $f(x;\varepsilon,\mu)$ by successive application of the limits ($\varepsilon \to 0$, $\overline{x},\overline{\mu}$ fixed). Thus

$$f_1 = \lim_{\varepsilon \to 0}\left(\frac{f}{\alpha_1}\right), \quad f_2 = \lim_{\varepsilon \to 0}\left(\frac{f-\alpha_1 f_1}{\alpha_2}\right), \ \ldots$$

These expansions are not usually valid in the entire region (a,b) but have some restricted domain of validity, in which the asymptotic nature of (1) remains. When the validity breaks down (non-uniformity) an adjacent expansion is sought, which can be matched to the previous one in an overlap domain.

In the following sections three examples are presented in which asymptotic analysis leads to a better understanding. In the first involving water waves there is a non-uniformity at infinity and also as the parameter Froude number F→1. The resolution of the non-uniformity at infinity is a non-linear far field expansion. The resolution of that as F→1 seems to demand a numerical calculation of an unsteady flow. The second example of transonic slender body theory shows a boundary value problem which is singular because the body shrinks to a line. The inner expansion preserves body geometry and is valid for small radius. The outer expansion preserves non-linear transonic effects and is valid from some region near the body to infinity. These expansions do not really match so that an intermediate expansion, valid for radius of order of the body length is necessary. The third example considers thin supersonic wings whose edges are swept to (or close to) the Mach angle. Although the linearized theory gives a finite answer it is not correct because of a local singularity at the edge. An expansion valid near the edge, containing non-linear transonic terms, must be constructed and matched to the linear solution away from the edge. An analogy is thus shown between the steady three-dimensional flow near the edge and two-dimensional unsteady small-disturbance transonic flow.

Another connection of asymptotic and numerics is the asymptotic analysis of finite difference schemes as the mesh-size approaches zero, and the analysis of truncation error. These aspects are not discussed here.

II. TWO DIMENSIONAL FREE SURFACE FLOW PAST A BUMP. The exact problem for an inviscid heavy fluid is illustrated in

Fig. 1. Continuity and irrotationality lead to the Laplace equation. Flow is assumed steady. Boundary conditions of tangent flow are applied at the free-surface and at the bump and the Bernoulli equation is used to express p=0 on the free surface.

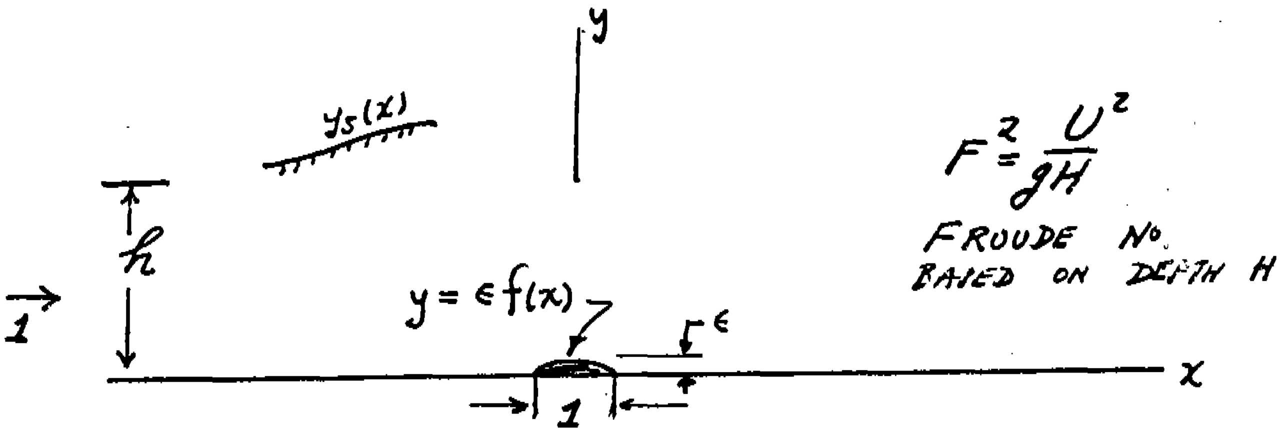


FIG. 1. EXACT PROBLEM

In the exact formulation velocities are scaled by the free-stream speed U, lengths by the bump length L, pressure by ρU^2. The exact problem is

$$\phi_{xx} + \phi_{yy} = 0 \qquad \text{(II-1)}$$

$$\left\{\begin{array}{l} \phi_y(x, y_s(x)) = y_s'(x)\phi_x(x, y_s(x)) \\ \frac{F^2h}{2}\left\{\phi_x{}^2 + \phi_y{}^2\right\} + y = \frac{F^2h}{2} + h \end{array}\right\} \quad \text{free surface} \qquad \text{(II-2)}$$

$$\begin{array}{ll} \phi_y(x, \epsilon f(x) = \epsilon f'(x)\phi_x(x, \epsilon f(x)) & |x| < \frac{1}{2} \\ \qquad = 0 & |x| > \frac{1}{2} \end{array} \quad \text{tangency} \qquad \text{(II-3)}$$

The parameters are

ϵ = thickness ratio of bump

F = Froude No. $= U/\sqrt{gH} = \dfrac{\text{flow speed}}{\text{fastest linear wave speed}}$

h = H/L .

For a small bump $\varepsilon<<1$ a limit process expansion associated with the limit $\varepsilon\to 0$, $(x,y),(F^2,h)$ fixed linearizes the problem. There is a small perturbation of uniform flow . The form for the potential ϕ, and free surface shape y_s is

$$\phi(x,y;\varepsilon,F,h) = x + \varepsilon\varphi(x,y;F,h) + \varepsilon^2\varphi_2(x,y;F,h) + \ldots$$

$$y_s(x;\varepsilon,F,h) = h + \varepsilon\eta(x;F,h) + \varepsilon^2\eta_2(x;F,h) + \ldots$$

The linear problem that results was solved by Kelvin and Lamb [2]

$$\varphi_{xx} + \varphi_{yy} = 0 \qquad \text{(II-5)}$$

$$F^2h\varphi_{xx}(x,h) + \varphi_y(x,h) = 0 \qquad \text{(II-6)}$$

$$\varphi_y(x,0) = f'(x) \quad . \qquad \text{(II-7)}$$

The solution has the features shown in Fig. 2.

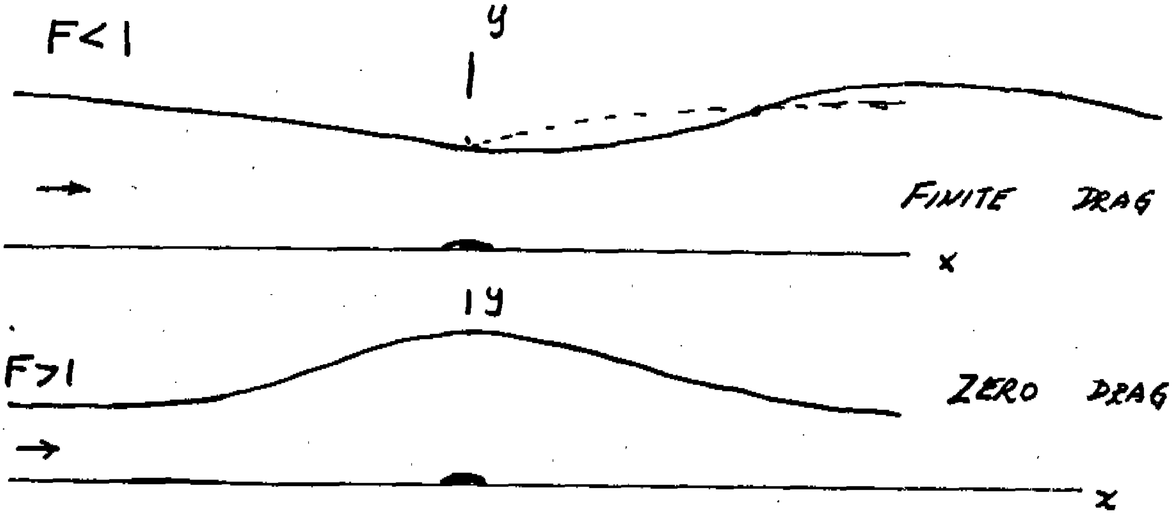


FIG. 2. LINEARIZED FLOW PAST BUMP

For F<1 the flow dips down over the bump and rises with a wave train behind. There is a finite wave drag. For F>1 no linear standing waves are possible. The flow rises over the bump and falls behind. The drag is zero. The linear expansion has a non-uniformity as $F\to 1-$;for F<1 the wave length gets longer as $F\to 1-$ and finally the solution has the asymptotic form (if η_2 is worked out)

$$\eta \to -\frac{3x}{h^2} \quad , \qquad \eta_2 \to -\frac{9}{4h^7}x^4 \qquad \text{(II-8)}$$

From this it is clear that there is a non-uniformity downstream when εx^3-1. Non-linear wave effects become important downstream.

In order to study this a new limit is considered in which F→1 and the observer goes farther and farther downstream as $\varepsilon \to 0$. That is

$$\varepsilon \to 0, \quad (\tilde{x}=\varepsilon^{1/3}x, y), (K,h) \quad \text{fixed where } F^2=1-K\varepsilon^{2/3} .$$

K is a similarity parameter. The powers of ε are chosen for a distinguished limit. x changes are slow. The Laplace equation is approximated by $\phi_{yy}=0$ and in effect, iterated. The form is

$$\phi = x + \varepsilon^{1/3}a_1(\tilde{x}) + \varepsilon\{a_2(\tilde{x})-y^2a_1''(\tilde{x})/2\}+\varepsilon^{5/3}\{a_3(\tilde{x})-y^2a_2''(\tilde{x})/2 + y^4a_1''''(\tilde{x})/24\} + \dots$$

$$y_s = h + \varepsilon^{2/3}\xi_1(\tilde{x}) + \varepsilon^{4/3}\xi_2(\tilde{x}) + \dots \qquad \text{(II-9)}$$

The equation for the free surface that holds downstream is

$$\frac{d^2\xi_1}{dx^2} + \frac{9}{2h^3}\xi_1^{\ 2} + \frac{3K}{h^2}\xi_1 = 0 \qquad \text{(II-10)}$$

All the possible solutions appear in the phase plane (θ,ξ_1) of Fig. 3.

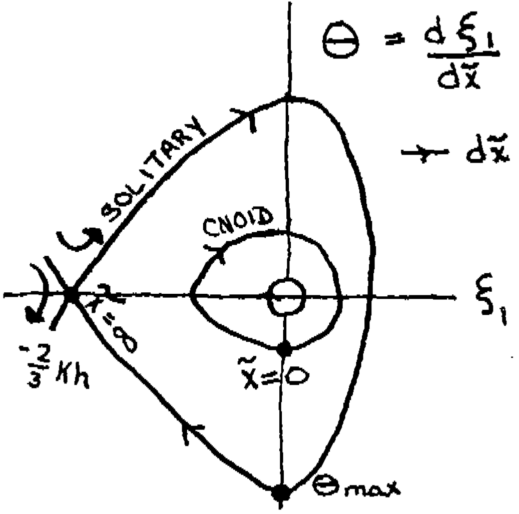


FIG. 3. PHASE PLANE

The solution is chosen so that, for matching, as $\tilde{x}\to 0+$, $\theta \to -3/h^2$ (cf II-8). A is related to the area of the bump. If $\theta(\tilde{x}=0)$ is of sufficiently small magnitude there is matching to a closed trajectory, a downstream non-linear (cnoidal) wave train periodic in $\tilde{x}$. For $K=K_{cr}$, θ appears at θ_{max} and then the solution is the downstream part of a solitary wave. The ultimate level downstream is lower than that upstream. But no solution exists for $K<K_{cr}$ so that there is a further non-uniformity as F gets closer to 1. The supercritical solution F>1 also blows up as F→1+. For a further discussion of this problem see [3].

III. TRANSONIC SLENDER BODY THEORY. Transonic slender body theory shows a non-uniformity of a different nature. Consider a slender body as in Fig. 4 flying close to the speed of sound. The theory is worked out in the framework of gas dynamics (inviscid, perfect gas). Shocks are weak so that the flow remains isentropic and a potential Φ exists. Lengths are scaled by the body length, velocities by the free-stream speed U. The small parameter δ is the body thickness ratio and measures the order of the flow deflection. The angle of attack α, not appearing explicitly, is assumed $O(\delta)$. The continuity equation becomes the compressible potential equation

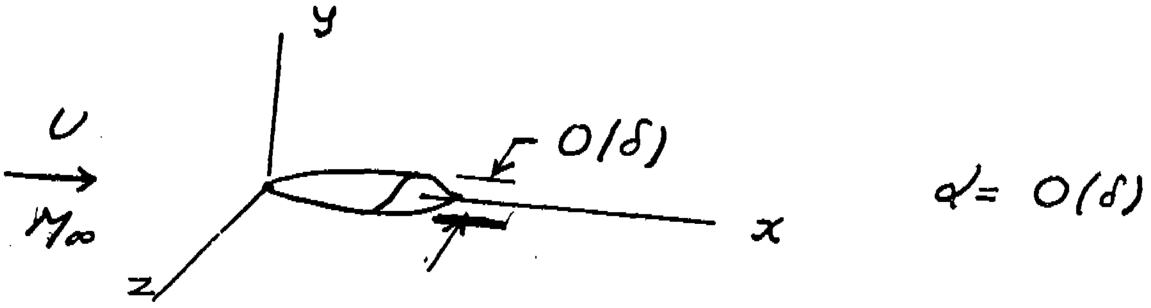


FIG. 4. SLENDER BODY

$$a^2\nabla^2\Phi = \vec{q}\cdot\nabla\left(\frac{q^2}{2}\right) \quad , \quad \vec{q} = \nabla\Phi \qquad \text{(III-1)}$$

where

$$\frac{a^2}{U^2} = \frac{1}{M_\infty^2} + \frac{\gamma-1}{2}\left\{1-\frac{q^2}{U^2}\right\} , \quad \gamma\text{=ratio of specific heats}=\frac{c_p}{c_v} \; . \qquad \text{(III-2)}$$

(III-2) is the compressible Bernoulli equation, a=local speed of sound, M_∞=free-stream Mach number=U/a_∞. Equation (III-1) is non-linear of changing type, locally elliptic where the flow is subsonic $|\vec{q}|<a$, and locally hyperbolic where the flow is supersonic $|\vec{q}|>a$. The boundary condition of tangent flow has to be satisfied on the body surface B=0,

$$\text{on } B\left(x,\frac{y}{\delta},\frac{z}{\delta}\right) = 0 \quad , \quad \vec{q}\cdot\nabla B = 0 \; . \qquad \text{(III-3)}$$

The approximations are concerned with $\delta\to 0$, $M_\infty\to 1$. In outer coordinates the body shrinks to a line. An inner expansion is needed to preserve the body geometry and satisfy (III-3). Thus the limit $\delta\to 0$, $(x, y^*=\frac{y}{\delta}, z^*=\frac{z}{\delta})$ fixed is considered. At the same time $M_\infty \to 1$ such that $K = \frac{1-M_\infty^2}{\delta^2}$ fixed, as considerations of the outer limit show is necessary. The inner expansion is of the form

$$\Phi = U\Big\{x+\delta^2\log\delta 2S(x)+\delta^2\varphi(x,r^*,\theta)+\delta^4\log^2\delta\varphi_{22} + \delta^4\log\delta\varphi_{21}+\delta^4\varphi_2+\ldots\Big\} \qquad \text{(III-4)}$$

$$\text{where } r^{*2}=y^{*2}+z^{*2} \quad , \ \theta=\tan^{-1}\left(\frac{y^*}{z^*}\right).$$

This expansion includes so-called switch back terms introduced for purposes of matching. The dominant equation of the inner expansion is the Laplace equation in each cross-section plane

$$\varphi_{y^*y^*}+\varphi_{z^*z^*} = 0 \qquad \text{(III-5)}$$

The outer expansion, valid to infinity has to preserve the non-linear changing type structure. Since disturbances spread laterally to much greater distances than upstream, the observer has to run to infinity laterally. The limit process has $\delta \to 0$ $(x,\tilde{r}=\delta r,\theta;K)$ fixed where $r=\sqrt{y^2+z^2}$. The expansion is of the form (anticipating some matching)

$$\Phi = U\{x+\delta^2\phi(x,\tilde{r})+\delta^4\phi_2(x,\tilde{r},\theta)+\ldots\} . \qquad \text{(III-6)}$$

The resulting dominant equation is the transonic small-disturbance equation

$$(K-(\gamma+1)\phi_x)\phi_{xx} + \phi_{\tilde{r}\tilde{r}} + \frac{1}{\tilde{r}}\phi_{\tilde{r}} = 0 \qquad \text{(III-7)}$$

Matching is concerned with behavior as $r^* \to \infty$, $\tilde{r} \to 0$. We know that far away the solution of (III-5) looks like a source-line, with $S(x)$ a source strength

$$\varphi = S(x)\log r^* + g(x) + \ldots \qquad \text{(III-8)}$$

Formally matching for the first terms shows that indeed ϕ is axisymmetric. But the two expansions do not really match as can be inferred from the fact that $r \to \infty$ in the outer limit and $r \to 0$ in the inner. The difficulty is manifest in the higher orders. An intermediate expansion is thus needed and turns out to be connected with the limit $\delta \to 0$, $(x,y,z;K)$ fixed. The form is

$$\Phi = U\{x+\delta^2\overline{\phi}(x,y,z;K)+\delta^3\overline{\phi}_1+\delta^4\overline{\phi}_2+\ldots\} . \qquad \text{(III-9)}$$

The equations are

$$\Phi yy + \bar{\phi}zz = 0 \qquad \text{(III-10)}$$

$$\bar{\phi}1_{yy} + \bar{\phi}1_{zz} = 0 \qquad \text{(III-11)}$$

$$\bar{\phi}_{2yy} + \bar{\phi}_{2zz} = (\gamma+1)\bar{\phi}_x\bar{\phi}_{xx} - K\bar{\phi}_{xx} + \frac{\partial}{\partial x}(\bar{\phi}_y{}^2+\bar{\phi}_z{}^2) \quad . \qquad \text{(III-12)}$$

The intermediate expansion must match to the inner expansion as $r=\sqrt{y^2+z^2} \to 0$, $(r^*\to\infty)$ and to the outer as $r\to\infty$ $(\tilde{r}\to 0)$. Matching shows that

$$\bar{\phi} = S(x)\log r + g(x) \quad . \qquad \text{(III-13)}$$

The first term on the right hand side of (III-12) is dominant as $r\to\infty$ and is the non-linear term necessary to match to the outer expansion as $\tilde{r}\to 0$, (III-7). The last term on the right hand side of (III-12) dominates as $r\to 0$ and is the non-linear term necessary to match to the higher order inner terms.

The outcome of all this is a well posed boundary value problem for (III-7) $r\phi_r \to S(x)$ $r\to 0$ where $S(x)$, from the inner expansion, is related to the rate of change for cross-section area of the slender body. The dominant outer flow is axisymmetric. This defines $g(x)$. Further a deeper understanding of the non-axisymmetric flow fields is gained.

IV. PSEUDO-TRANSONIC FLOW. In the same general framework as the preceeding section, pseudo-transonic effects occur when the leading edge of a thin supersonic ($M_\infty > 1$) wing is swept to the Mach angle $\theta_M = \tan^{-1} \frac{a_\infty}{U}$. (See Fig. 5)

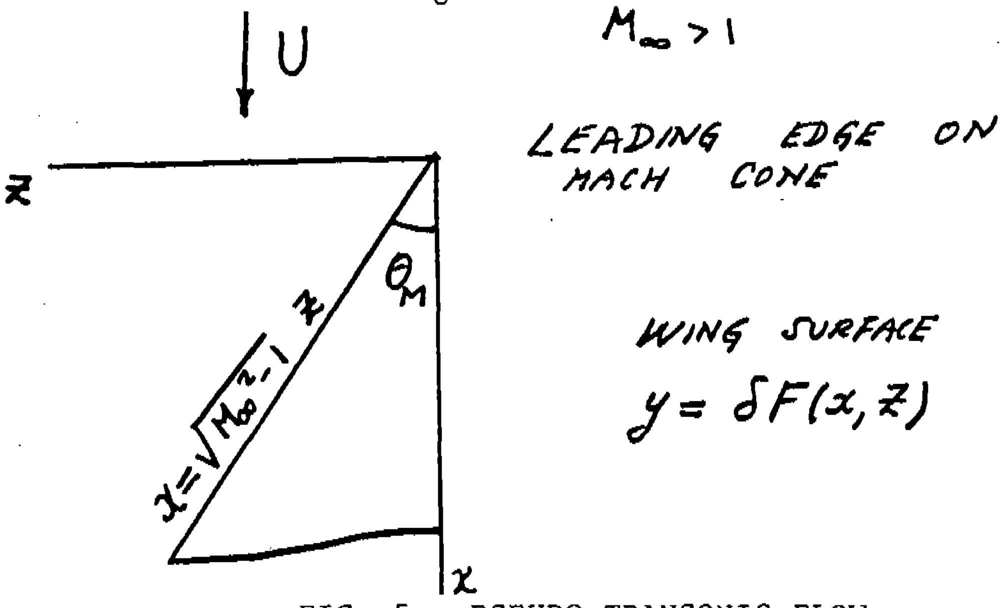


FIG. 5. PSEUDO-TRANSONIC FLOW

δ now measures the thickness ratio of the wing. Linearized supersonic theory is traditionally used to calculate the flow about such wings. The physical content is that of acoustics. The limit process has $\delta \to 0$ $(x,y,z;M_\infty)$ fixed. The corresponding expansion is of the form

$$\Phi = U\{x+\delta\phi(x,y,z;M_\infty)+\ldots\} \qquad \text{(IV-1)}$$

The disturbance potential ϕ satisfies the classical wave equation

$$(M_\infty{}^2-1)\phi_{xx} - (\phi_{yy}+\phi_{zz}) = 0 \qquad \text{(IV-2)}$$

The boundary value problem in $y>0$ can be solved by a distribution of supersonic sources over the wing planform in $y=0$. For the special case of a wedge airfoil $F(x,z) = (x-\sqrt{M_\infty{}^2-1}\;z)\frac{1}{2}$ the boundary condition of tangent flow is

$$\phi_y(x,0+,z) = \frac{1}{2} \qquad \text{(IV-3)}$$

The solution is

$$\phi(x,y,z) = -\frac{1}{\pi}\iint_{\sqrt{\ }\text{real}} \frac{(1/2)d\xi d\zeta}{\sqrt{(x-\xi)^2-\beta^2(z-\zeta)^2-\beta^2y^2}}\ , \quad \beta^2=M_\infty{}^2-1 \qquad \text{(IV-4)}$$

$$= -\frac{1}{\pi\beta}\sqrt{x^2-\beta^2(y^2+z^2)} + \frac{y}{\pi}\cos^{-1}\frac{\beta y}{\sqrt{x^2-\beta^2z^2}} \qquad \text{(IV-5)}$$

The square root behavior indicates a singularity in ϕ_x at the leading edge and a non-uniformity of linearized theory. The component of flow normal to the edge is exactly sonic so that transonic effects could be expected. Although the singularity is integrable the finite results in this case are not in good agreement with experiment. A better treatment locally comes from the following limit

$$\delta \to 0 \quad \left(\xi = \frac{x-\beta z}{\delta^{2/3}}\ ,\ \eta = \frac{y}{\delta^{1/3}}\ ,\ \zeta=z\ ,\ M_\infty\right) \text{ fixed.}$$

The expansion is valid in a local region near the leading edge. The form turns out to be

$$\Phi = U\{x+\delta^{4/3}\ (\xi,\eta,\zeta;M_\infty)+\ldots\} \qquad \text{(IV-6)}$$

The distinguished equation for φ that results is

$$(\gamma+1)M_\infty{}^4\varphi_\xi\varphi_{\xi\xi} - \varphi_{\eta\eta} + 2\beta\varphi_{\xi\zeta} = 0 \qquad \text{(IV-7)}$$

This equation is in perfect analogy with the unsteady version of the two-dimensional transonic small disturbance equation (III-7) with K=0. Here $\xi \to x$, $\eta \to \tilde{y}$, $\zeta \to \tilde{t}$. This says that the flow can be computed as if there were unsteady flow at each spanwise section starting at the apex. As usual (IV-7) admits shock waves. The solution far away from the edge must match to the linear solution (IV-5).If we write (IV-5) in inner coordinates and keep dominant terms we see that

$$\phi \to -\frac{1}{\beta\pi}\sqrt{\delta^{2/3}2\beta\xi\zeta-\delta^{2/3}\beta^2\eta^2}+\delta^{1/3}\frac{\eta}{\pi}\cos^{-1}\frac{\delta^{1/3}\beta\eta}{\sqrt{\delta^{2/3}2\beta\xi\zeta}}$$

Thus for matching

$$\varphi \to -\frac{1}{\beta\pi}\sqrt{2\beta\xi\zeta-\beta^2\eta^2}+\frac{\eta}{\pi}\cos^{-1}\frac{\beta\eta}{\sqrt{2\beta\xi\zeta}} \qquad \text{as } \xi,\eta\to\infty \qquad \text{(IV-8)}$$

This provides the far field boundary condition for (IV-7), and calculations can be performed. This particular problem is conical and can be simplified further. It can be shown that the drag including pseudotransonic effects is reduced.

The results here indicate that the flow near a high aspect ratio wing swept to the Mach angle should also be approximated by (IV-7). This is the case for the wing in Fig. 6.

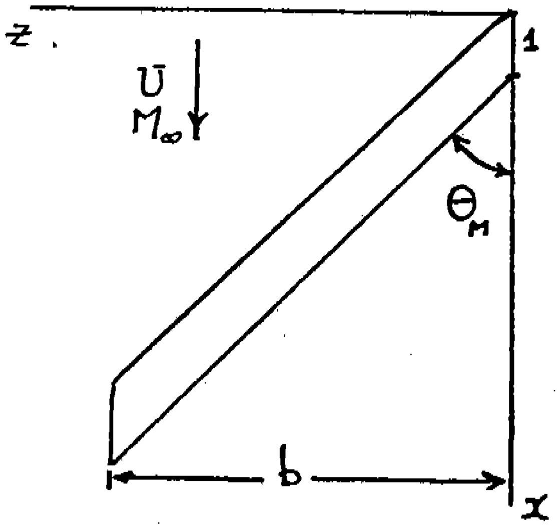


FIG. 6. SWEPT HIGH ASPECT RATIO WING

The expansion is of the form

$$\Phi(x,y,z) = U\{x+\delta^{2/3}\ (\xi,\eta,\zeta;B+,\ldots\} \tag{IV-9}$$

where $\xi=x-\beta z$, $\eta=\delta^{1/3}y$, $\zeta=\delta^{1/3}z$ and the similarity parameter $B=\delta^{1/3}b$.

The results of this section are easily extended to cases where the edges are close the Mach angle.

ACKNOWLEDGMENT: The work of this author was supported in part by Air Force Contract AFOSR-82-0155.

REFERENCES

1. J. Kevorkian and J.D. Cole., Perturbation Methods in Applied Mathematics", Springer-Verlag, New York 1981.

2. H. Lamb, Hydrodynamics, Dover Press.

3. S.L. Cole, "Near Critical Free Surface Flow Past an Obstacle", Quarterly of Applied Mathematics, To appear.

AN EQUIVALENT GAS MODEL FOR DUSTY GASES

Donald A. Drew* and Fredrick J. Zeigler**

1. INTRODUCTION

Many fluid flow problems of interest concern the behavior of a gas which has been contaminated with small particles of dust. The presence of the dust can cause significant changes in the flow, and it is important to analyze an explanation for this phenomenon. This is done by examining a model in which the gas and dust exchange heat and momentum. In the limit of low volumetric concentrations of dust, but with strong coupling between the phases, the model equations are closely approximated by the equations for an adiabatic ideal gas, with modified values of the density and ratio of specific heats. By the use of similarity transformations of these equations, it is possible to relate solutions of flow problems for a gas with dust to solutions of corresponding problems for a clear gas, thus giving an explicit way of calculating the effect of the dust on the flow. Because of their simple form, the equations of transonic flow are used to provide an example of this procedure. It is found that the transonic flow around a thin airfoil for a gas with dust is equivalent to the flow around an airfoil with modified thickness, at a different free-stream Mach number.

*Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12181
**General Dynamics Corporation, Fort Worth Division, P. O. Box 748, Mail Zone 2882, Ft. Worth, TX 76101

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

2. EQUATIONS OF TWO-PHASE FLOW

The equations governing the flow of particles and fluid are

$$\frac{\partial \alpha\rho_p}{\partial t} + \nabla\cdot\alpha\rho_p\vec{q}_p = 0 \ , \tag{2.1a}$$

$$\frac{\partial(1 - \alpha)\rho_f}{\partial t} + \nabla\cdot(1 - \alpha)\rho_f\vec{q}_f = 0 \ , \tag{2.1b}$$

$$\alpha\rho_p\left(\frac{\partial\vec{q}_p}{\partial t} + \vec{q}_p\cdot\nabla\vec{q}_p\right) = -\alpha\nabla p + \alpha b_M(\vec{q}_f - \vec{q}_p) \ , \tag{2.2a}$$

$$(1 - \alpha)\rho_f\left(\frac{\partial\vec{q}_f}{\partial t} + \vec{q}_f\cdot\nabla\vec{q}_f\right) = -(1 - \alpha)\nabla p + \alpha b_M(\vec{q}_p - \vec{q}_f) \ , \tag{2.2b}$$

$$\alpha\rho_p\left[\frac{\partial\left(\varepsilon_p + \frac{1}{2}q_p^2\right)}{\partial t} + \vec{q}_p\cdot\nabla\left(\varepsilon_p + \frac{1}{2}q_p^2\right)\right] =$$

$$= -\nabla\cdot\alpha p\vec{q}_p - p\nabla\alpha\cdot\vec{q}_p + \alpha b_M(\vec{q}_f - \vec{q}_p)\cdot\vec{q}_p + \alpha H_M(T_f - T_p) \ , \tag{2.3a}$$

$$(1 - \alpha)\rho_f\left[\frac{\partial\left(\varepsilon_f + \frac{1}{2}q_f^2\right)}{\partial t} + \vec{q}_f\cdot\nabla(\varepsilon_f + \frac{1}{2}q_f^2)\right] =$$

$$= -\nabla\cdot(1 - \alpha)p\vec{q}_f - p\nabla\alpha\cdot\vec{q}_p - \alpha b_M(\vec{q}_f - \vec{q}_p)\cdot\vec{q}_p + \alpha H_M(T_p - T_f) \ , \tag{2.3b}$$

where 2.1, 2.2, and 2.3 are equations of conservation of mass, momentum, and energy for the particle and fluid phases. Here $\vec{q}$ represents the velocity (p for particles, f for fluid), α is the dust volumetric density, ρ denotes the phasic density, p is the pressure, and ε is the internal energy per unit mass.

The dust is assumed to be incompressible

$$\rho_p = \text{const} \ ; \tag{2.4}$$

the ideal gas law

$$p = R\rho_f T_f \tag{2.5}$$

holds for the fluid. We assume

$$\varepsilon_p = c_v^{(p)} T_p ,$$

and,

$$\varepsilon_f = c_v^{(f)} T_f + p/\rho_f . \tag{2.6}$$

The terms on the right of (2.2) and (2.3) involving b_M and H_M reflect momentum and energy transfer between the particles and gas. For b_M we assume Stokes drag

$$b_M = \frac{9}{2} \rho_f \nu_f \frac{1}{a^2} . \tag{2.7}$$

The heat transfer coefficient H_M is assumed to be

$$H_M = \frac{1}{3} \text{Nu Pr } b_M . \tag{2.8}$$

where Nu is the Nusselt number and Pr is the Prandtl number.

The dusty gas limit of the previous two-phase flow equations corresponds to the limit $\alpha << 1$, $\rho_f << \rho_p$ and $\alpha\rho_p \sim \rho_f$. In order to exhibit this limit explicitly we first scale the variables and put the equations into nondimensional form.

Thus, we assume the problem contains a typical length scale L, velocity scale U, gas density scale Γ, temperature scale T_0, and volumetric concentration scale A. If we denote a dimensionless variable by a caret (^), we have

$$\frac{\partial\hat{\alpha}}{\partial\hat{t}} + \hat{\nabla}\cdot(\hat{\alpha}\hat{\vec{q}}_p) = 0 \tag{3.1a}$$

$$\frac{\partial\hat{\rho}_f}{\partial\hat{t}} + \hat{\nabla}\cdot\hat{\rho}_f\hat{\vec{q}}_f = 0 \tag{3.1b}$$

$$\frac{\partial\hat{\vec{q}}_p}{\partial\hat{t}} + \hat{\vec{q}}_p\cdot\hat{\nabla}\hat{\vec{q}}_p = \frac{1}{\varepsilon f}\,(\hat{\vec{q}}_f - \hat{\vec{q}}_p) \tag{3.2a}$$

$$\hat{\rho}_f\Big(\frac{\partial\hat{\vec{q}}_f}{\partial\hat{t}} + \hat{\vec{q}}_f\cdot\hat{\nabla}\hat{\vec{q}}_f\Big) = -\hat{\nabla}\hat{p} + \frac{1}{\varepsilon}\,\hat{\alpha}(\hat{\vec{q}}_p - \hat{\vec{q}}_f) \tag{3.2b}$$

$$\Big[\frac{\partial(\hat{\varepsilon}_p + \frac{1}{2}\,\hat{q}_p^2)}{\partial\hat{t}} + \hat{\vec{q}}_p\cdot\hat{\nabla}(\hat{\varepsilon}_p + \frac{1}{2}\,\hat{q}_p^2)\Big] = \frac{1}{\varepsilon f}\,(\hat{\vec{q}}_f - \hat{\vec{q}}_p)\cdot\hat{\vec{q}}_p + \frac{h}{\varepsilon f}\,(\hat{T}_f - \hat{T}_p) \tag{3.3a}$$

$$\hat{\rho}_f\Big[\frac{\partial(\hat{\varepsilon}_f + \frac{1}{2}\,q_f^2)}{\partial\hat{t}} + \hat{\vec{q}}_f\cdot\hat{\nabla}(\hat{\varepsilon}_f + \frac{1}{2}\,\hat{q}_f^2)\Big] =$$

$$= -\hat{\nabla}\cdot\hat{p}\hat{\vec{q}}_f - \frac{1}{\varepsilon}\,\hat{\alpha}(\hat{\vec{q}}_f - \hat{\vec{q}}_p)\cdot\hat{\vec{q}}_p + \frac{h}{\varepsilon}\,\hat{\alpha}(\hat{T}_p - \hat{T}_f)\ . \tag{3.3b}$$

The equations of state take the form

$$\hat{p} = r\hat{\rho}_f\hat{T}_f\ , \tag{3.4}$$

$$\hat{\varepsilon}_p = c\hat{T}_p\ , \tag{3.5}$$

$$\hat{\varepsilon}_f = \hat{c}\hat{T}_f + \hat{p}/\hat{\rho}_f\ . \tag{3.6}$$

The dimensionless constants are defined by

$$\varepsilon = \left(\frac{Ua}{\nu_f}\right)\left(\frac{a}{L}\right)\left(\frac{1}{A}\right)\left(\frac{2}{9\hat{\rho}_f}\right) . \tag{3.7a}$$

$$h = H_M T_0/b_M U^2 \tag{3.7b}$$

$$r = RT_0/U^2 \tag{3.3c}$$

$$c = c_v^{(p)} T_0/U^2 \tag{3.3d}$$

$$\hat{c} = c_v^{(f)} T_0/U^2 . \tag{3.3e}$$

4. GENERALIZED GAS MODEL

We see that for sufficiently small particles, ε will be small. The rest of our discussion will be limited to the regime $\varepsilon << 1$.

For this case, inspection of equation 3.2 shows that $\vec{q}_f \approx \vec{q}_p$ (we now drop the carets for convenience) unless the accelerations are large. This suggest a model which we term the generalized gas model, which has the property that $\vec{q}_p = \vec{q}_f + 0(\varepsilon)$, except in places where the flow fields change rapidly. We shall call these regions of rapid change generalized shocks. As we shall see, this model is analogous to a gas with changed properties.

Consider the flow fields α, $\vec{q}_p$, $\vec{q}_\rho$, etc. as functions of $\vec{x}$, t, and ε. In the generalized gas model, we consider a limit of the dusty gas equations in which $\varepsilon \to 0$, with $\vec{x}$, t held fixed. Away from generalized shocks, we therefore consider an expansion

$$\alpha(x,t;\varepsilon) = \alpha^{(0)}(x,t) + \varepsilon\alpha^{(1)}(x,t) + \ldots \tag{4.2}$$

with similar expressions for the other flow quantities.

Substituting these expressions into the dusty gas equations and equating terms of equal order to zero gives the following:

$$\vec{q}_f^{(0)} = \vec{q}_p^{(0)} \equiv \vec{q} \tag{4.3}$$

$$T_f^{(0)} = T_p^{(0)} \equiv T \tag{4.4}$$

$$\frac{\partial\alpha^{(0)}}{\partial t} + \nabla\cdot\alpha^{(0)}\vec{q} = 0 \tag{4.5}$$

$$\frac{\partial\rho_f^{(0)}}{\partial t} + \nabla\cdot\rho_f^{(0)}\vec{q} = 0 \tag{4.6}$$

$$f\left(\frac{\partial\vec{q}}{\partial t} + \vec{q}\cdot\nabla\vec{q}\right) = \vec{q}_f^{(1)} - \vec{q}_p^{(1)} \tag{4.7}$$

$$\rho_f^{(0)}\left(\frac{\partial\vec{q}}{\partial t} + \vec{q}\cdot\nabla\vec{q}\right) = -\nabla p^{(0)} + \alpha^{(0)}(\vec{q}_p^{(1)} - \vec{q}_f^{(1)}) \tag{4.8}$$

$$f\left[\frac{\partial(cT + \frac{1}{2}q^2)}{\partial t} + \vec{q}\cdot\nabla(cT + \frac{1}{2}q^2)\right] = (\vec{q}_f^{(1)} - \vec{q}_p^{(1)})\cdot\vec{q} + h(T_f^{(1)} - T_p^{(1)}) \tag{4.9}$$

$$\rho_f^{(0)}\left[\frac{\partial(\varepsilon_f^{(0)} + \frac{1}{2}q^2)}{\partial t} + \vec{q}\cdot\nabla(\varepsilon_f^{(0)} + \frac{1}{2}q^2)\right] =$$

$$= -\nabla\cdot p^{(0)}\vec{q} - \alpha^{(0)}(\vec{q}_f^{(1)} - \vec{q}_p^{(1)})\cdot\vec{q} + h\alpha^{(0)}(T_p^{(1)} - T_f^{(1)}) \tag{4.10}$$

$$p^{(0)} = r\rho_f^{(0)}T \tag{4.11}$$

$$\varepsilon_f^{(0)} = \hat{c}T + p^{(0)}/\rho_f^{(0)} \ . \tag{4.12}$$

This model allows us to derive the equations needed at the lowest order. Adding $\alpha^{(0)}$ times (4.7) to (4.8) yields

$$(\rho_f^{(0)} + \alpha^{(0)}f)\left(\frac{\partial\vec{q}}{\partial t} + \vec{q}\cdot\nabla\vec{q}\right) = -\nabla p^{(0)} \tag{4.13}$$

A similar combination of (4.9) and (4.10) gives

$$\rho_f^{(0)}\left[\frac{\partial\hat{c}T}{\partial t} + \vec{q}\cdot\nabla\hat{c}T\right] + \alpha^{(0)}f\left[\frac{\partial cT}{\partial t} + \vec{q}\cdot\nabla cT\right] +$$

$$+ (\rho_f^{(0)} + \alpha^{(0)}f)\left[\frac{\partial\left(\frac{1}{2}q^2\right)}{\partial t} + \vec{q}\cdot\nabla\left(\frac{1}{2}q^2\right)\right] +$$

$$+ \rho_f^{(0)}\left[\frac{\partial(p^{(0)}/\rho_f^{(0)})}{\partial t} + \vec{q}\cdot\nabla(p^{(0)}/\rho_f^{(0)})\right] = -\nabla\cdot p^{(0)}\vec{q} \ . \tag{4.14}$$

Additionally, (4.5) and (4.6) may be combined as

$$\frac{\partial(\rho_f^{(0)} + \alpha^{(0)}f)}{\partial t} + \nabla\cdot(\rho_f^{(0)} + \alpha^{(0)}f)\vec{q} = 0 = \frac{\partial\rho_m}{\partial t} + \nabla\cdot\rho_m\vec{q} \tag{4.15}$$

where $\rho_m = \rho_f^{(0)} + \alpha^{(0)}f$ is the mixture density. Using (4.5) and (4.6) we obtain

$$\frac{\partial}{\partial t}\, \ell n\left(\frac{\alpha^{(0)}}{\rho_f^{(0)}}\right) + \vec{q}\cdot\nabla \ell n(\alpha^{(0)}/\rho_f^{(0)}) = 0 \ . \tag{4.16}$$

Therefore $\alpha^{(0)}/\rho_f^{(0)}$ is constant for a fluid particle. We shall assume that $\alpha^{(0)}/\rho_f^{(0)}$ is constant over the entire flow domain. If we subtract the kinetic energy equation from (4.14), and use (4.6) and (4.11), we have

$$(\rho_f^{(0)}\hat{c} + \alpha^{(0)}fc + \rho_f^{(0)}r)\left[\frac{\partial T}{\partial t} + \vec{q}\cdot\nabla T\right] = rT\left[\frac{\partial \rho_f^{(0)}}{\partial t} + \vec{q}\cdot\nabla\rho_f^{(0)}\right] \tag{4.17}$$

Therefore

$$T = \text{Const}\cdot(\rho_f^{(0)})^{\hat{\gamma}-1} \tag{4.18}$$

where

$$\hat{\gamma} = \frac{\hat{c} + \alpha^{(0)}fc/\rho_f^{(0)}}{\hat{c} + r + \alpha^{(0)}fc/\rho_f^{(0)}} = \frac{\gamma + \dfrac{\alpha^{(0)}f}{\rho_f^{(0)}}\,\dfrac{c}{\hat{c} + r}}{1 + \dfrac{\alpha^{(0)}f}{\rho_f^{(0)}}\,\dfrac{c}{\hat{c} + r}} \tag{4.19}$$

Dropping the (0) superscript for the pressure, the generalized gas model is

$$\frac{\partial \rho_m}{\partial t} + \nabla\cdot\rho_m\vec{q} = 0 \tag{4.20}$$

$$\rho_m\left(\frac{\partial\vec{q}}{\partial t} + \vec{q}\cdot\nabla\vec{q}\right) = -\nabla p \ , \tag{4.21}$$

$$p/p_0 = (\rho_m/\rho_0)^{\hat{\gamma}} \ , \tag{4.22}$$

where ρ_0 and p_0 are constants. Thus, a dusty gas with small dust particles behaves like a gas with a modified γ. We emphasize that this derivation assumes that we are not near a shock.

We note that the speed of sound for a generalized gas is given by

$$a^2 = \frac{dp}{d\rho_m} = \frac{\hat{\gamma} p}{\rho_m} = \frac{\hat{\gamma}\rho_f^{(0)}}{\gamma\rho_m}\left(\frac{\gamma p}{\rho_f^{(0)}}\right) = \frac{\hat{\gamma}\rho_f^{(0)}}{\gamma\rho_m}\,\hat{a}^2 \; ,$$

where $\hat{a}$ is the speed of sound in the clear gas. Moreover, since $\rho_m = \rho_f^{(0)} + \alpha^{(0)} f$, $\dfrac{\gamma\rho_f^{(0)}}{\gamma\rho_m} = \dfrac{\hat{\gamma}}{\gamma}\,\dfrac{1}{1 + \dfrac{\alpha^{(0)} f}{\rho_f^{(0)}}}$.

5. SMALL DISTRUBANCE THEORY

The dusty gas may be regarded as an equivalent gas with a modified value of γ. Assuming that we have potential flow, it is therefore possible to formulate a small-disturbance theory for transonic thin airfoils in dusty gases, analogous to that in [2]. The transonic similarity parameter for this case is a function of the new value of γ. It will be shown how this modified similarity parameter is related to the usual one for the case of no dust, so that a method of estimating the effect of the dust on the usual small-disturbance theory may be achieved.

We formulate the boundary value problem for a thin airfoil, in a dusty gas and travelling in the transonic range. The free-stream velocity is U (in the x direction), where the coordinates x, y have been normalized with respect to the airfoil chord. The airfoil is given by $y = \delta F_{u,\ell}(x)$, where the function F satisfies $\max\limits_{x \in [0,1]} |F_u(x) - F_\ell(x)| = 1$. The free stream Mach number is $\hat{M}_\infty = U/\hat{a}_\infty$.

Under the condition of small disturbances, we may derive an approximate equation by a limit process expansion for the velocity potential Φ, based on the limit process $\delta \to 0$, with $x, \tilde{y} = \delta^{1/3} y$, and $\hat{K} = \dfrac{1 - \hat{M}_\infty^2}{\delta^{2/3}}$ fixed as $\hat{M}_\infty \to 1$. It has the form

$$\Phi(x,y;\delta) = U\{x + \delta^{2/3}\phi(x,\tilde{y}) + \ldots\} \ . \tag{5.1}$$

Then ϕ satisfies

$$(\hat{K} - (\hat{\gamma} + 1)\phi_x)\phi_{xx} + \phi_{\tilde{y}\tilde{y}} = 0 \tag{5.2}$$

and the boundary conditions

$$\phi_{\tilde{y}}(x,0^\pm) = F'_{u,\ell}(x) \qquad 0 < x < 1 \tag{5.3}$$

$$\phi_x, \phi_{\tilde{y}} \to 0 \quad \text{as} \quad x \to -\infty \ . \tag{5.4}$$

In this context, the Kutta condition may be written as

$$[\phi_x]_{TE} = 0 \ , \tag{5.5}$$

where TE means trailing edge.

Additionally, shock jump conditions must be imposed in order to have a complete problem for ϕ.

The form of (5.2) shows that for two flows at different transonic Mach numbers, but with the same values of $\hat{K}$, these two flows will be geometrically similar (the difference will be that the 'size' of the disturbance will be determined by different factors of $\delta^{2/3}$). An analogous similarity law holds for the gas alone, governed by the similarity parameter K. Both of these similarity laws relate families of flows, for different values of the displacement thickness at a corresponding Mach number.

We wish to see how the change in the ratio of the specific heats for the dusty gas, $\hat{\gamma}$, affects the flow. This will be done by finding a correspondence between the similarity parameters in the two cases.

We shall develop the correspondence by transforming the boundary value problem (5.2)-(5.5) for the dusty gas into a problem for the gas without dust. Define

$$\bar{K} = \omega^{-2}\hat{K} \tag{5.6a}$$

$$\bar{y} = \omega\tilde{y} = \bar{\delta}^{1/3} y \tag{5.6b}$$

$$\bar{\phi} = \omega\phi \tag{5.6c}$$

where $\omega = \left(\frac{\hat{\gamma} + 1}{\gamma + 1}\right)^{1/3}$, and $\bar{\delta} = \omega^3\delta$. Then, under (5.6) the entire problem for $\phi(x,\tilde{y};\hat{K})$ transforms to the corresponding problem for $\bar{\phi}(x,\bar{y};\bar{K})$, which is the case of the gas without dust. The similarity parameter $\bar{K}$ for the new problem is of the form

$$\bar{K} = \frac{1 - \bar{M}_\infty^{-2}}{\bar{\delta}^{2/3}} = \frac{1 - \bar{M}_\infty^{-2}}{(\omega^3 \delta)^{2/3}} = \frac{1 - \bar{M}_\infty^{-2}}{\omega^2 \delta^{2/3}} = \omega^{-2}\hat{K} \ , \tag{5.7}$$

from which we conclude $\bar{M}_\infty = \hat{M}_\infty$.

Thus, we see from (5.6c) that the small disturbance problem for a dusty gas is equivalent to one for a clear gas, with a modified amplitude of disturbances, a changed wing thickness ratio and similarity parameter.

The transformation parameter ω, as a function of $\alpha^{(0)}f/\rho_f^{(0)}$, is shown in Figure 5.1 for $c/(\hat{c} + r) = 1.0$, which is a representative value for many materials.

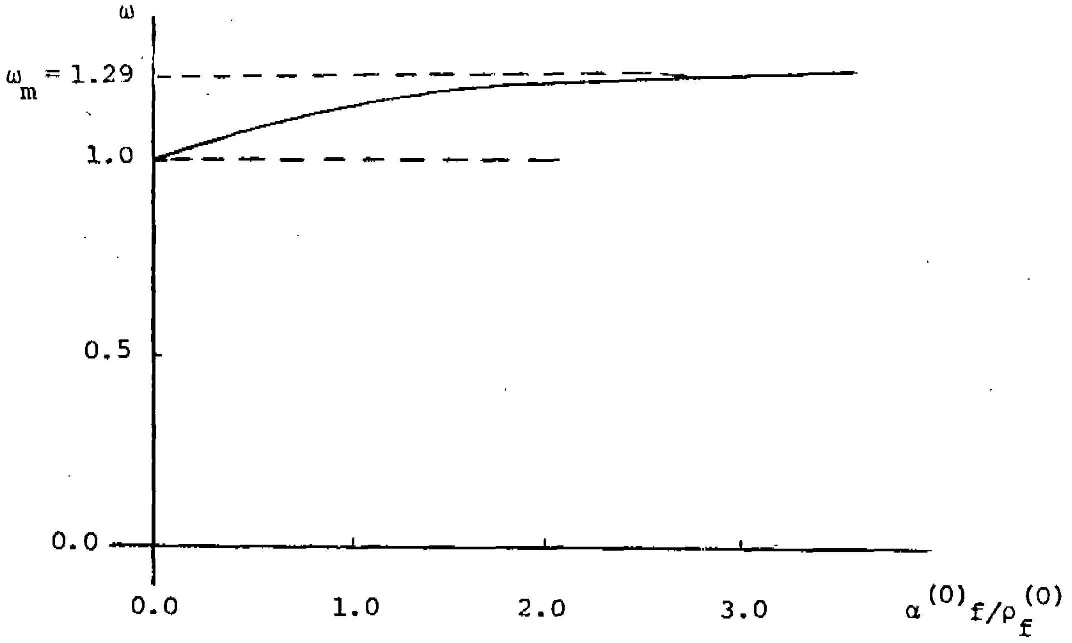


Figure 5.1. ω vs. $\alpha^{(0)}f/\rho_f^{(0)}$

Thus, for example, a dusty flow with a mass loading of 2.0 corresponds to a value of ω of 1.21, which means that if the airfoil has thickness δ, the flow is equivalent to that of a clear gas around an airfoil of the same shape, but of thickness $\bar{\delta} = \omega^3\delta = 1.77\delta$. The equivalent similarity parameter $\bar{K}$ is $\bar{K} = \omega^{-2}\hat{K} = 0.68\hat{K}$.

REFERENCES

1. Ishii, M., 1975, Thermo-fluid dynamic theory of two-phase flow, Eyrolles, Paris, France.
2. J. Kevorkian and J. D. Cole, Perturbation Methods in Applied Mathematics, Chapter 5, Springer-Verlag, 1981.
3. Bolz, R. E. and Tuve, G. L., eds., 1973, Handbook of tables for applied engineering science, 2nd ed., Chemical Rubber Co., Cleveland, OH.

INSTABILITY OF THE FLOW OF TWO IMMISCIBLE LIQUIDS WITH DIFFERENT VISCOSITIES IN A PIPE

Daniel D. Joseph*, Michael Renardy and Yuriko Renardy
Mathematics Research Center
University of Wisconsin - Madison
610 Walnut Street
Madison, WI 53705

ABSTRACT. We study the flow of two immiscible fluids of different viscosities and equal density through a pipe under a pressure gradient. This problem has a continuum of solutions corresponding to arbitrarily prescribed interface shapes. The question therefore arises, which of these solutions are stable and thus observable. Experiments have shown a tendency for the thinner fluid to encapsulate the thicker one. This has been "explained" by the viscous dissipation principle, which postulates that the amount of viscous dissipation is minimized for a given flow rate. For a circular pipe, this predicts a concentric configuration with the more viscous fluid located at the core. A linear stability analysis, which is carried out numerically, shows that while this configuration is stable when the more viscous fluid occupies most of the pipe, it is not stable when there is more of the thin fluid. Therefore the dissipation principle does not always hold, and the volume ratio is a crucial factor.

I. INTRODUCTION.

The flow we consider is a cylindrical pipe of infinite length in which there are two fluids. The fluids are immiscible and have the same density but different viscosities. The flow is steady, purely axial and driven by a prescribed pressure gradient.

The equations governing the flow are the steady Navier-Stokes equations with the velocity and the pressure gradient in the axial direction, and incompressibility. The boundary conditions are: no slip at the pipe wall, and at the interface of the two fluids, which one of the unknowns, the normal and shear stresses and the velocity are to be continuous. We specify the ratio of the cross-sectional area occupied by each fluid and ask the question: what shape will the interface be?

Theoretically, it is known that if there is no surface tension, every interface position is allowed by the equations. If there is surface tension, then the interface has to be circles or circular arcs terminating at the pipe wall. The number of possible steady solutions is still infinite. For example, if you specify that fluid 1 occupies 1/3 of the area and fluid 2 occupies 2/3 of the area, 2 possible arrangements are shown in Figure 1.

* Department of Aerospace Engineering and Mechanics, 107 Akerman Hall, University of Minnesota, Minneapolis, MN 55444.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

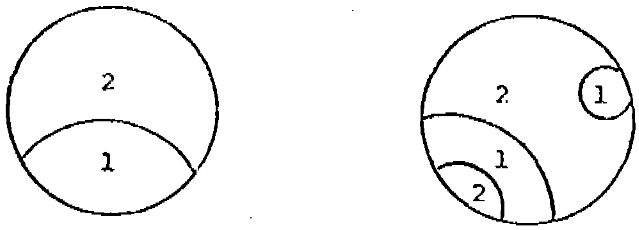


Figure 1

Such nonuniqueness appears in the theory of steady 2-fluid flows for all kinds of flow regimes [1]. On the other hand, experiments with the pipe flow indicate that whatever the initial configuration, the low viscosity liquid will eventually encapsulate the thicker fluid. The encapsulation property has been observed for both high and low Reynolds number flows, ranging from oil and water to molten polymers [2, 3, 4, 5, 6, 7, 8, 9].

What has to be done is to reconcile the existence of a continuum of solutions with the experimentally observed unique configuration. Up to now, explanations have been based on a variational method, called the "viscous dissipation principle" which says that the flow chooses an interface which in some sense minimizes viscous dissipation for a given flow rate, or equivalently, maximizes the volume flux for a given pressure gradient [2, 5, 8, 10]. This is based on the idea that there is work to be done, but it is harder to make the thick fluid do the work, so the thin fluid does it by migrating to regions of high shear. The thin fluid is easy to push around.

Michael Renardy has shown that for a pipe with arbitrary cross-section, the minimizer of viscous dissipation exists if there is an a priori estimate for the length of the interface curve. This is a quantity which we do not know how to obtain [11]. Not much more is known about the interface for flows in pipes of arbitrary cross-section. However, for a circular pipe, the analysis is much simpler because of the symmetry and the minimizer turns out to be the concentric configuration with the thick fluid at the core. This appears to agree with experiments. Our question is: how valid is the viscous dissipation principle? One way to find out is to do a stability analysis for the circular pipe to see if the configuration preferred by the viscous dissipation principle turns out to be stable.

II. NUMERICAL CALCULATIONS. We did a linear stability analysis for the circular pipe where the basic flow is the Poiseuille flow with a concentric interface (Figure 2). Fluid 1 is at the core, fluid 2 encapsulates fluid 1. We superimpose an infinitesimal disturbance $(u,v,w,p)e^{i(-\alpha ct+\alpha z+n\theta)}$. We use a Chebyschev polynomial expansion in the radial direction [12]. The problem is then an eigenvalue problem for c, given all the other parameters. If the sign of the imaginary part of c is positive, the flow is unstable to small disturbances.

The particular case of the long-wave limit, $\alpha \times$ Reynolds number $\to 0$, and the thinner fluid at the core was studies by Hickox [13] and was shown to be unstable. This supports the viscous dissipation principle, but Hickox did not look at the case where the thicker fluid is at the core to see if that would be stable.

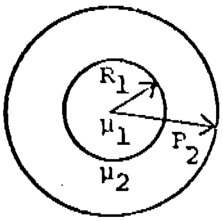


Figure 2

III. RESULTS. The eigenvalue which determines instability is an interfacial one in the sense that it is neutrally stable when the two viscosities are equal. This situation is not identical to the one-fluid flow because of the extra conditions at the interface. Yih [14] found similar results when he looked at plane Couette flow with a flat interface with the long-wave approximation.

Our range of parameters is the following: viscosity ratio μ_1/μ_2 from 0.2 to 8, dimensionless wavelength of axial disturbance αR_2 from 0.1 to 10, reference Reynolds number $R_2V(R_1)/\nu_2$ from 0 to 1000, where $V(R_1)$ is the basic velocity at the interface and ν_2 is the kinematic viscosity, density is taken to be 1.

First, we found that the configuration with the thin fluid at the core is unstable. This extends Hickox's long-wave results and agrees with the viscous dissipation principle. Secondly, when the thick fluid is at the core, stability depends on the radius ratio R_1/R_2. This is what we found to be most interesting because it shows that the viscous dissipation principle is not always true. This dependence of the stability on the radius ratio is qualitatively similar to Yih's [14] results where stability depends on the depth ratio of the 2 fluids. Figure 3 shows an example of what we found at Reynolds number 100, $\alpha R_2 = 1$.

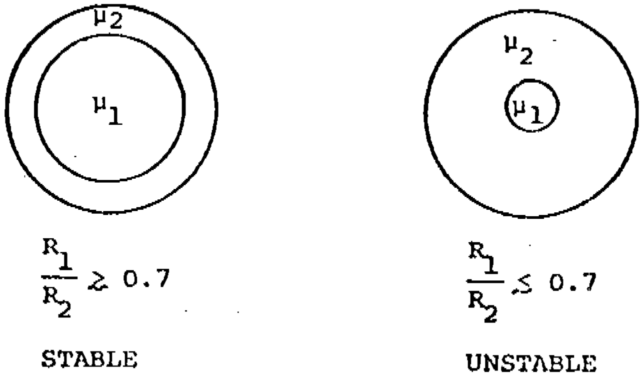


Figure 3

$$\frac{\mu_1}{\mu_2} > 1$$

Figure 4

Re = 100, $\alpha R_2 = 1$, $R_1/R_2 = 0.7$

The magnification on the right displays behaviour for $1 \le \frac{\mu_1}{\mu_2} \le 2$.

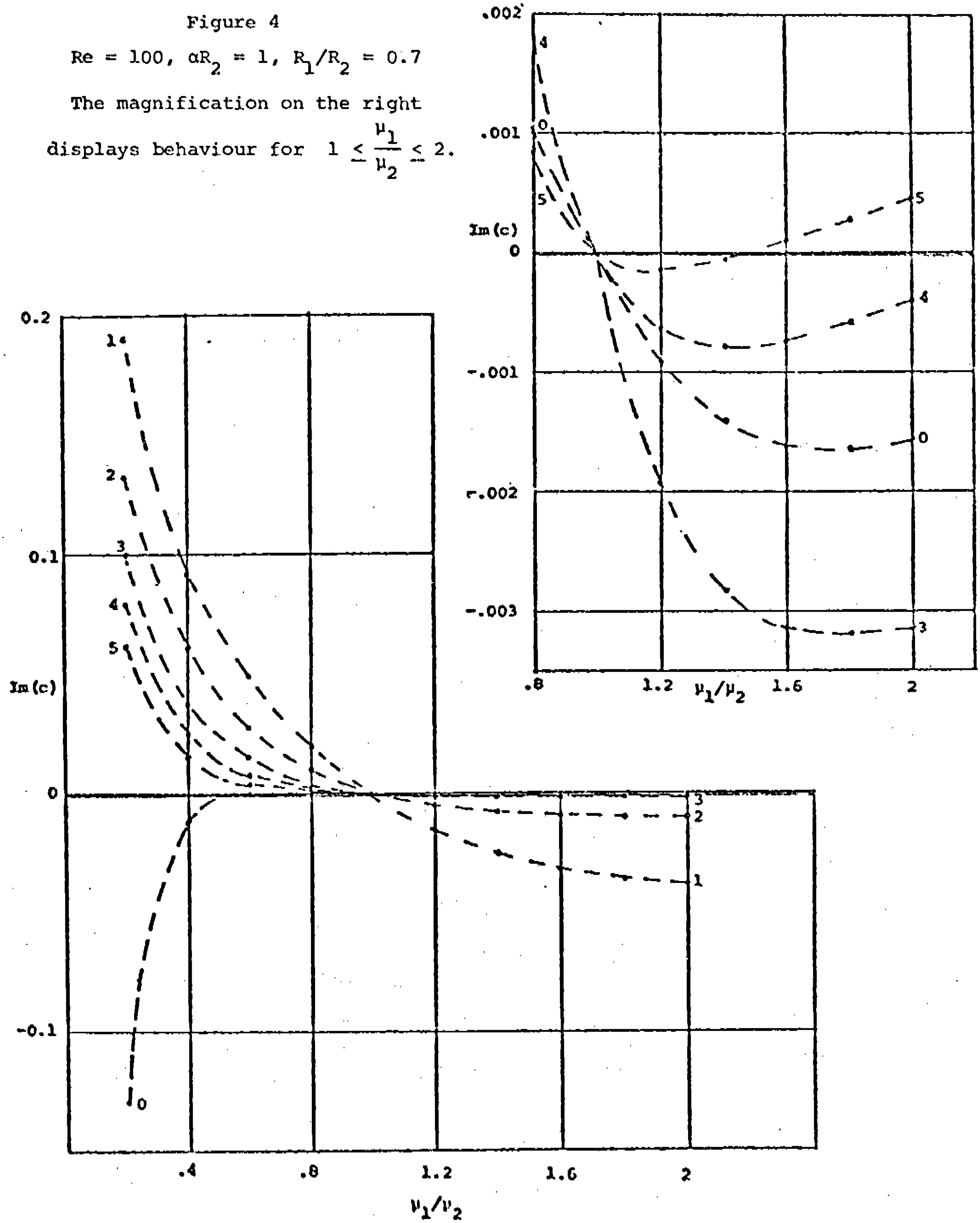


Figure 4 is a graph of the imaginary part of c versus viscosity ratio for Re = 100, $\alpha R_2 = 1$, $R_1/R_2 = 0.7$. Numbers next to the curves denote azimuthal mode numbers. The dark points on the curves show our computed values and the dashed lines are interpolants. At this radius ratio, there is a slight instability due to the higher azimuthal modes. (Mode 5 becomes positive in the inset.) For $\frac{R_1}{R_2} \gtrsim 0.7$, the curves sink below the Im(c) = 0 line for $\frac{\mu_1}{\mu_2} > 1$. For $\frac{R_1}{R_2} \lesssim 0.7$, the curves rise above the line, yielding the results in Figure 3.

Thirdly, when the viscosity ratio is large, the response changes only gradually. Figure 5 is a graph at $\frac{R_1}{R_2} = 0.9$, Re = 100, $\alpha R_2 = 1$. Whether $\frac{\mu_1}{\mu_2}$ is 6 or 7, it does not make much difference to the imaginary part of c. This behaviour has been mentioned in some experiments [2].

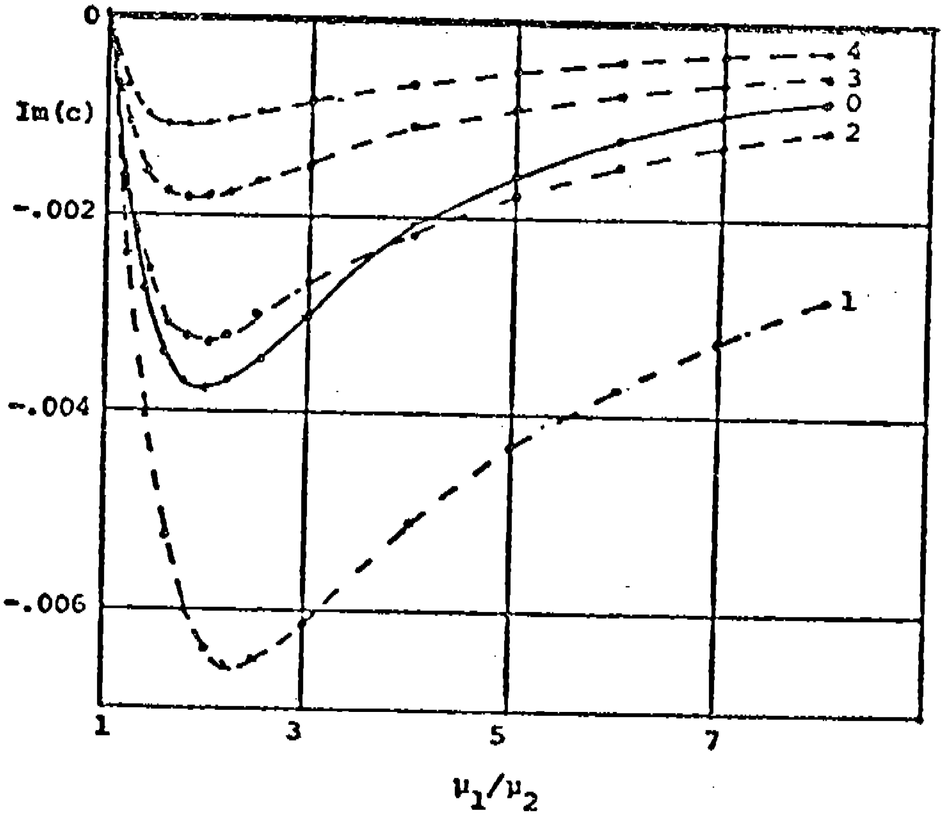


Figure 5

Re = 100, $\alpha R_2 = 1$, $R_1/R_2 = 0.9$

Fourthly, for short disturbance wavelengths, stability is lost. We did not include surface tension in our computations and that would dampen some of these instabilities. Figure 6 is a graph of im(c) versus viscosity ratio at $Re = 100$, $\frac{R_1}{R_2} = 0.8$, $\alpha R_2 = 10$. When $\alpha R_2 = 1$, the region of stability is $\frac{\mu_1}{\mu_2} > 1$ but for $\alpha R_2 = 10$, the modes are mostly unstable. This agrees with the results of Hooper and Boyd [15] who consider the linear stability of an unbounded Couette flow. The 2 fluids occupy each half-plane. Their analysis is relevant locally at any interface with a viscosity jump and predicts instability for short-wave disturbances. This is in contrast with one-fluid flows where viscosity acts to dampen short waves.

Fifthly, as the Reynolds number increases, stability is lost. Figure 7 is a graph of $Re = 1000$, $\frac{R_1}{R_2} = 0.8$, $\alpha R_2 = 1$. When $Re = 100$, the region of stability is $\frac{\mu_1}{\mu_2} > 1$ but when $Re = 1000$ the 0^{th} mode is unstable. The region of stability at $Re = 1000$ for Figure 7 is for $\frac{\mu_1}{\mu_2} \gtrsim 1.8$.

IV. CONCLUSION. Our conclusion is that the viscous dissipation principle is not always true, but that does not mean we are saying it is never true. From our results (Figure 3), there seems to be some truth to the basic idea that the thin fluid tends to lubricate the wall. However, if there is a large enough amount of the thin fluid, other mechanisms must be at work. Also, the situation we have dealt with has a rotational symmetry and symmetric solutions to symmetric problems always have a special status so that it is quite natural for the concentric configuration to be preferred over others in this case, whatever the mechanisms may be.

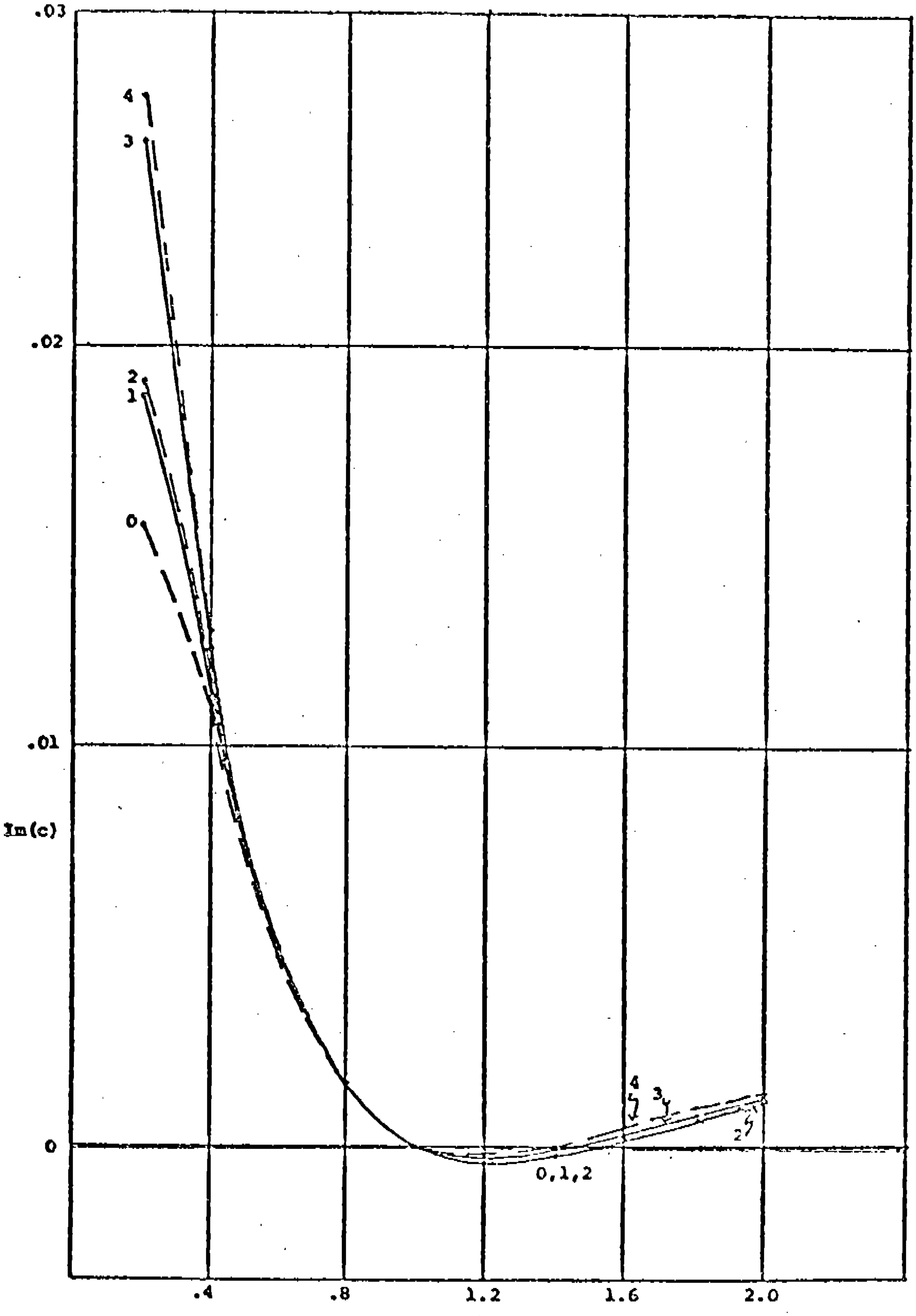


Figure 6

Re = 100, αR_2 = 10, R_1/R_2 = 0.8

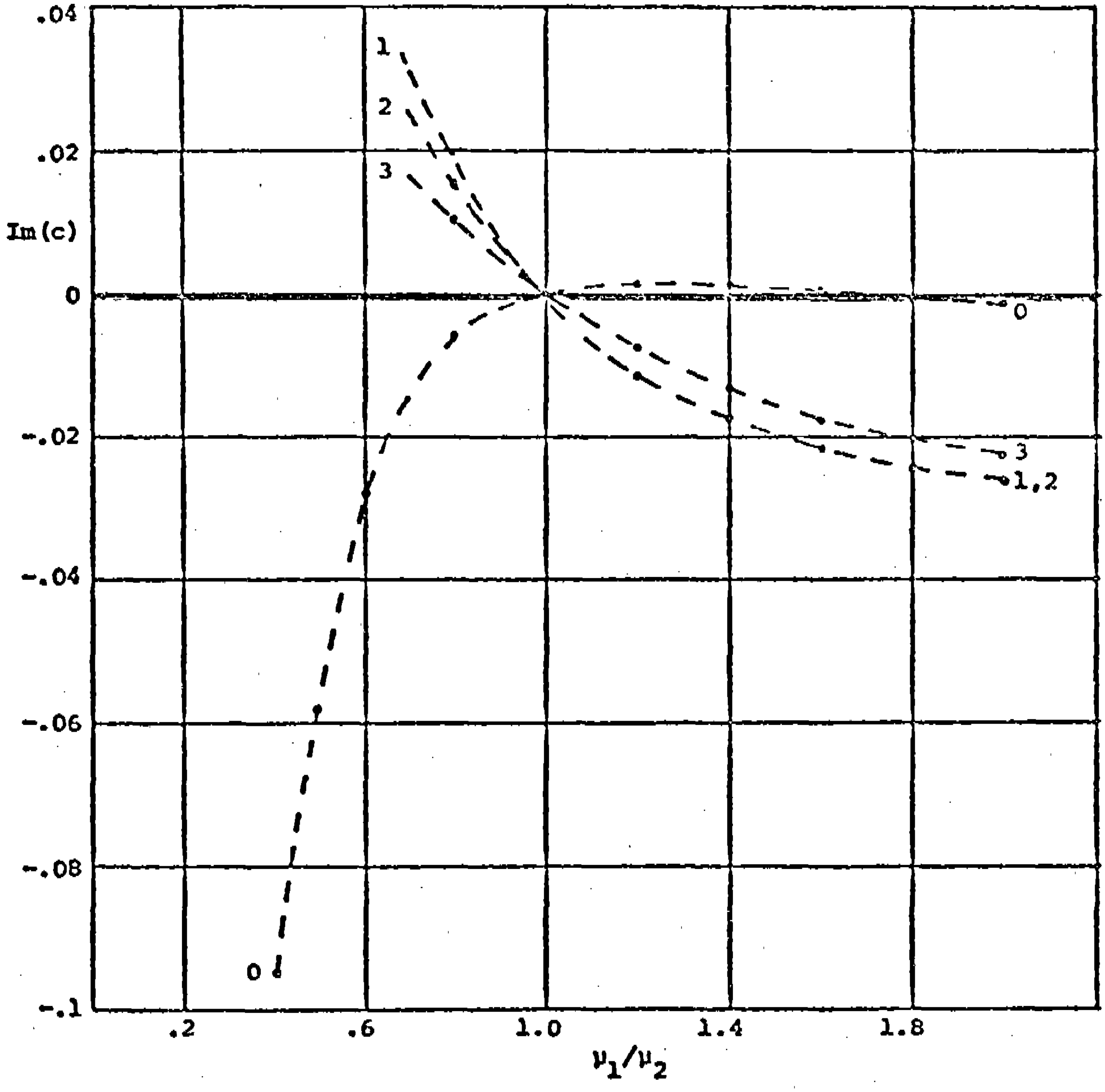


Figure 7

Imaginary part of c versus viscosity ratio
Re = 1000, αR_2 = 1, R_1/R_2 = 0.8, velocity scale = 1

REFERENCES

[1] Joseph, D. D., Nguyen, K. and Beavers, G. S., Nonuniqueness and Stability of the Configuration of Flow of Immiscible Fluids with Different Viscosities, to appear.

[2] Everage, A. E. Jr., Theory of Bicomponent Flow of Polymer Melts. I. Equilibrium Newtonian Tube Flow, Trans. Soc. Rheology 17:4, 629-646 (1973).

[3] Hasson, D. and Nir, A., Annular Flow of Two Immiscible Liquids. II. Analysis of Core-Liquid Ascent, Canadian J. Chem. Eng. 48, 521-526 (1970).

[4] Minagawa, N. and White, J. L., Co-Extrusion of Unfilled and TiO_2-Filled Polyethylene: Influence of Viscosity and Die Cross-Section on Interface Shape. Polymer Engineering and Science, 15, 825-830 (1975).

[5] Southern, J. H. and Ballman, R. L., Stratified Bicomponent Flow of Polymer Melts in a Tube. Appl. Polymer Symp. No. 20, 175-189 (1973).

[6] White, J. L. and Lee, Biing-Lin, Theory of Interface Distortion in Stratified Two-Phase Flow. Trans. Soc. Rheology, 19:3, 457-479 (1975).

[7] Charles, M. E. and Redberger, R. J., The Reduction of Pressure Gradients in Oil Pipelines by the Addition of Water. Numerical Analysis of Stratified Flows. Canadian J. Che. Eng. 40, 70-75 (1962).

[8] Williams, M. C., Migration of Two Liquid Phases in Capillary Extrusion: An Energy Interpretation, AICHE Journal, 21, 1204-1207 (1975).

[9] Yu, H. S. and Sparrow, E. M., Experiments on Two-Component Stratified Flow in a Horizontal Duct. J. Heat Transfer, 91, 51-58 (1969).

[10] MacLean, D. L., A Theoretical Analysis of Bicomponent Flow and the Problem of Interface Shape. Trans. Soc. Rheol. 17:3, 385-399 (1975).

[11] Joseph, D. D., Renardy, M. and Renardy, Y., Instability of the Flow of Immiscible Liquids With Different Visconsities in a Pipe. MRC Technical Summary Report #2503.

[12] Orszag, S. A. and Kells, L. C., Transition to Turbulence in Plane Poiseuille and Plane Couette Flow, J. Fluid Mech. 96, 159-205 (1980).

[13] Hickox, Charles E., Instability Due to Viscosity and Density Stratification in Axisymmetric Pipe Flow, Physics of Fluids, 14, 251-262 (1971).

[14] Yih, C. S., Instability Due to Viscosity Stratification, J. Fluid Mech. 27, 337-352 (1967).

[15] Hooper, A. P. and Boyd, W. G. C., Shear-flow Instability at the Interface Between Two Viscous Fluids, J. Fluid Mech., 128, 507-528 (1983).

NUMERICAL SIMULATION OF THE FLOW OF A CHEMICAL CONTAMINANT SUBJECTED TO A JET IMPINGEMENT

L. M. Chang
Interior Ballistics Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. A numerical simulation and study are presented for characterization of the interaction flow of a water jet with a chemical contaminant droplet on a plane wall, which occurs in chemical decontamination processes. The model for this analysis is a two-dimensional, two-fluid flow governed by the Navier-Stokes equations. Emphases of the study are on the evolution of the contaminant droplet and the effects of variation of relevant flow parameters on the displacement of the contaminant. Computed results show that a jet impingement at an angle of incidence in the range of 45° - 60° can achieve the highest cleaning speed of the contaminant. The results also show that an increase in either the jet velocity or its cross-sectional area can greatly improve the cleaning speed. However, for a given jet flow rate, it is more advantageous to increase the jet velocity rather than the cross-sectional area in order to increase the efficiency of the use of jet fluid.

I. INTRODUCTION. Application of jet spray for removal of chemical contaminants of solid walls has long been recognized to be effective. The procedure is to use the great force produced by the turning of the jet stream to displace the contaminant which is in the form of droplets along the walls.

A high-performance jet system for this use should possess the following basic feasures: high cleaning speed and efficient use of jet fluid. These two features are particularly important when the system is operated in the field where it is often required to decontaminate an area in the shortest period of time and with the least consumption of jet fluid. In designing such a system, knowledge of fundamental characteristics of the flow is vital, such as the evolution of the contaminant droplet and the effects of variation of relevant flow parameters on the jet performance.

The contaminant droplet has an average size of 3 mm in diameter and 0.6 mm in height. Its density is approximately the same as that of plain water, however, its viscosity may vary widely from 10 to 1000 times the viscosity of plain water.

The interaction flow between the jet and the contaminant droplet consists of two fluids, the jet fluid (water) and the contaminant, or three fluids if the ambient is treated as the third one. The two prime fluids are separated by interfaces and have free surfaces with the ambient. The flow is three-dimensional in nature and is highly transient. Most of the investigations conducted in the area of jet impingement in the past are relevant to the VTOL program (vertical takeoff and landing aircraft) or rocket exhaust flows, and are concerned with impingements on a solid surface [1,2,3,4,5,6]. For impingements on a liquid surface, Hund [7] and Vanden-Broeck [8] considered steady and two-dimensional cases. They used simplified theories to characterize the wave-like hydrodynamic instability occuring at the interface of the two fluids, but gave no predictions on the velocity and pressure distributions in the flow field.

In the present investigation, we have simplified the analysis by treating the jet-contaminant flow as a two-dimensional problem. We have developed a two-fluid flow model suitable for characterization of the flow and adoption of the existing computer code SOLA-VOF [9] for numerical solutions. The flow is governed by the unsteady Navier-Stokes equations. Presented are the evolution of the contaminant droplet and effects of variation of relevant flow parameters on the jet performance in displacing the droplet.

II. FLOW MODEL. Figure 1 depicts a pre-impingement flow configuration which practically occurs in the decontamination process. A water jet is directed at a contaminant droplet which is initially covered by a thin water layer on a plane wall. To characterize the flow developing from this configuration following the impingement we have developed a two-dimensional, two-fluid viscous flow model in Figure 2.

The flow region in the model is essentially the region enclosed by the dashed line indicated in Figure 1, covering the major part of the flow field. It resembles a channel flow containing two fluids (water and contaminant) separated by an interface. Both fluids are assumed to be newtonian and the surface tension effect along the interface is neglected because of its small magnitude. The upper boundary of the channel coincides with the upper free surface of the water layer so as to eliminate consideration of the free surface interface with the ambient. An outflow condition is specified at this boundary and at the ends of the channel, allowing the fluids to flow out of the region. The contaminant which initially occupies the shaded rectangular region is assumed to wet perfectly the bottom wall of the channel. To account for viscous effects, a no-slip condition is used for the bottom wall. Finally, a steady uniform jet velocity at an angle of θ is specified along a segment of the upper boundary as shown in the figure. The lower left side of the Figure 2 shows part of the computational domain of the flow field. The domain is discreted into a 100 x 12 mesh with finer zoning near the lower wall to ensure better accuracy of computation in this thin layer.

III. FLOW EQUATIONS AND METHOD OF SOLUTION. The governing equations for the model flow are:

continuity
$$\frac{1}{\rho c^2}\frac{\partial p}{\partial t} + \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \qquad (1)$$

momentum
$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right] \qquad (2)$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial y} + \nu\left[\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right] \qquad (3)$$

where t is time variable, u and v are the x-component (along the channel) and the y-component (normal to the channel) of the flow velocity, respectively.

The density ρ, the sound speed c, and the viscosity ν, are constant. In addition, a function F, called the fractional volume of fluid function, is introduced for tracking the water-contaminant interface. The function is given as

$$\frac{\partial F}{\partial t} + u \frac{\partial F}{\partial x} + v \frac{\partial F}{\partial y} = 0 \qquad (4)$$

This equation states that F moves with the fluid. In a two-fluid flow the value of F is unity at any point occupied by the first fluid (say, contaminant) and zero elsewhere. When averaged over the cells of a computational mesh, the average value of F in a cell is equal to the fractional volume of the cell occupied by the first fluid. In particular, a unity value of F corresponds to a cell full of the first fluid, whereas a zero value indicates that the cell contains only the second fluid (say, water including the jet fluid and the water layer of Figure 1). Cells with F values between zero and one contain an interface, as illustrated in Figure 3. With this, the interface separating the two fluids can be tracked.

The velocity components u and v in the momentum Eqs. (2) and (3) have been solved by using the explicit finite difference scheme, while the pressure p has been computed, coupled with the continuity equation (Eq. (1)), via an implicit finite difference method. The solution of the F function in Eq. (4) has been obtained by using the Donor-Accepter flux approximation. Details of the solution method have been given in Reference 9 of this paper. In order to observe the evolution (location and shape) of the region covered by the contaminant droplet, Marker Particles have been embedded in the fluid and move with it, but do not affect the fluid dynamics.

Numerical computations have been carried out by employing the SOLA-VOF code (9). In the current version of the code, the viscosities of both fluids in a flow are considered the same or simply zero. To adapt this code for solving the present flow which involves two fluids with very different viscosities we have implemented the following viscosity relationship into the code.

$$\nu = \nu_c F + (1 - F) \nu_w \qquad (5)$$

where ν is the kinematic viscosity of fluid in a cell, ν_c the kinematic viscosity of the first fluid (contaminant), ν_w the kinematic viscosity of the second fluid (jet fluid), and F the function defined in Eq. (4). Similarly, the density in a cell is approximated to be

$$\rho = \rho_c F + (1 - F) \rho_w \qquad (6)$$

where ρ_c and ρ_w are the density of the contaminant and plain water, respectively. From Eqs. (5) and (6), we see that the values of ν and ρ in a cell are functions of F.

Finally, it is noted that the Reynolds numbers based on the jet thickness and the jet velocities used in our computations are in the range of 20 - 2000. Within this range, Eqs. (2) and (3) are felt to be appropriate for the present flow analysis, even though the equations do not include turbulence considerations.

IV. COMPUTATIONAL RESULTS AND DISCUSSIONS. The following are the input data for our computations:

θ = incidence angle of the jet = 19.8° - 90°

D_j = diameter of the jet (thickness of the jet in the present two- dimensional model) = 1.83 mm

V_j = jet velocity, uniform across the jet = 5 - 10 m/sec

ρ_w = density of plain water = 0.001 Kg/cm^3

ρ_c = density of the contaminant = 0.00107 Kg/cm^3

ν_w = Kinematic viscosity of water = 0.0098 cm^2/sec

ν_c = kinematic viscosity of the contaminant = 0.098 - 9.8 cm^2/sec

Droplet size: 3 mm x 0.6 mm (length x height)

Computed results are presented as follows:

A. Flow Patterns. Figure 4 presents two typical flow developments following the commencement of the jet flow up to 0.2 millisecond. They correspond to the contaminant viscosities ν_c = 0.098 cm^2/sec (= 10 ν_w) and ν_c= 9.8 cm^2/sec (= 1000 ν_w), respectively. It is noted that all of the graphs in the figure have been magnified by a factor of 3 in the vertical direction in order to provide a better flow visualization near the bottom wall of the channel. The vectors indicated in the flow region represent the local fluid velocities in the individual cells of the computational mesh. In the low viscosity case the jet stream toward the droplet is very much parallel to the bottom wall, while in the high viscosity case the jet stream is lifted off the wall. It is interesting to note that as the time progresses, part of the fluid near the downstream edge of the droplet, especially in the high viscosity case, moves toward the bottom wall behind the droplet. Apparently, a low pressure region is created behind the lower portion of the droplet, similar to the flow region appearing behind an obstacle which stands in a flow. Figure 5 presents another view of the revolution of the two droplets. The viscosity effect is pronounced as seen when comparing the profiles and the displacements S of the two droplets. We also observe that at t = 0.2 millisecond all fluid particles inside the low viscosity droplet have been disturbed, while in the high viscosity droplet there is a region near the wall in which fluid particles remain in good order.

B. Optimum Jet Incidence Angle. The incidence angle of the jet impingement should play a key role in affecting the jet performance in displacing the contaminant droplet. A small incidence angle will result in a large portion of the jet stream moving toward the contaminant droplet, but a small normal force to displace the contaminant near the wall as a result of a small rate change of momentum in the impingement area. A large incidence angle, on the other hand, will produce a reverse result. Therefore, there is an optimum angle at which a jet impingement can perform best for removal of the contaminant.

In order to optimize the incidence angle, several angles ranging from 19.8° to 90° have been used to compute the displacement of the droplet as a function of time following the impingement process. At each angle the jet is properly located such that the impingement will need the shortest time to displace the upstream edge of the droplet a prescribed distance, say, one third of the original droplet length.

Figures 6 and 7 show the results for the jet velocities of 5 m/sec and 10 m/sec respectively. The dashed lines represent the results for the case that the contaminant viscosity is equal to 10 times viscosity of plain water ($\nu c = 10\ \nu_w$) and the solid lines for $\nu_c = 1000\ \nu_w$. As seen, the displacement S at a given time t increases with the incidence angle θ until θ reaches 56.28°. Beyond that angle, S falls except in the case that $V_j = 5$ m/sec and $\nu_c = 1000\ \nu_w$ shown in Figure 6. When the water layer which initially covers the droplet as shown in Figure 3 is increased from 0.2 mm to 0.6 mm, the result shown in Figure 8 also indicates that $\theta = 56.28°$ results in a better jet performance than any other smaller angle. An important conclusion then can be drawn that for a given jet velocity and a given jet size (i.e., the cross-sectional area of the jet) a jet impinging at an angle around 56° can achieve the highest cleaning speed of the contaminant.

The efficiency of the use of jet fluid, as well as the cleaning speed, is of concern since in some areas, especially in the battlefield, the supply of the jet fluid could be very limited. Thus we need a system capable of decontaminating a surface area with the least consumption of jet fluid.

In Figure 9, A denotes the surface area which has been decontaminated and V^* the volume of jet fluid consumed for the decontamination of the area. Then A/V^* is called the nominal area cleaned per unit volume of jet fluid consumed and thus can represent the efficiency of the use of the jet fluid. The figure explicitly shows that a jet at an angle of incidence between 45° and 60° from the impingement surface will achieve the most efficient use of jet fluid.

C. Improvement of Jet Performance by Increasing the Jet Velocity or the Jet Size. The jet velocity and its cross-sectional area are another two important flow parameters governing the performance of the jet impingement for decontamination applications.

In Figure 10 a comparison between curve "a" and curve "b" or between curve "$\bar{a}$ " and curve "$\bar{b}$ " indicates that an increase in the jet velocity V_j from 7.27 m/sec to 10 m/sec accelerates the cleaning speed substantially. Similar results can be obtained when the jet size D_j (i.e., the thickness of the jet in the two-dimensional case) is enlarged as seen from a comparison between curve "a^*" and curve "b^*" or between curve "$\bar{a}^*$ " and curve "$\bar{b}^*$ ". A question then arises: Which is the better means to improve the cleaning speed provided that the jet flow rate ($\dot{Q} = V_j D_j$) is the same, by increasing the jet velocity or increasing the jet size. For the answer, we first compare curve "a" with curve "a^*". As listed at the right corner of the figure, both curves correspond to the same jet flow rate $Q = 0.0183$ m^3/sec, however, curve "a" resulted from a jet velocity twice that of curve "a^*" but from a jet size half that of curve "a^*". The comparision between them reveals that the one (curve "a") resulted from a higher jet velocity provides a higher cleaning speed. This is also true when we compare curve "$\bar{a}$ " with curve "$\bar{a}^*$ ", curve "b" with curve "b^*", or curve "$\bar{b}$ " with curve $\bar{b}^*$ ".

Next we examine the increases of the jet velocity and the jet size in relation to the efficiency of the use of jet fluid. In Figure 11 the upper curve of each set of the curves is the result from an increase in jet velocity V_j from 5 m/sec to 10 m/sec, while the jet size is held constant, $D_j = d = 1.83$ mm. The lower curve of each set is the result from an increase in jet size from d to 2d, while the jet velocity remains the same, $V_j = 5$ m/sec. We see that the efficiency of the use of jet fluid, represented by A/V^*, increases with the increasing jet velocity, but decreases with the increasing jet size. The figure shows that at a given jet flow rate $\dot{Q}$, the jet with an increased jet velocity performs more efficiently in terms of consuming less jet fluid to decontaminate a given area.

As a remark, the above results obtained from the present two-dimensional (2D) flow model are also applicable to the three-dimensional (3D) impingement. Unlike in the 2D impingement in which most of the jet fluid from the increased jet size flows toward the contaminant, only a small amount of it moves toward the contaminant in the 3D case. Accordingly, in the 3D impingement an increase in the jet size is less effective than in the jet velocity in improving the jet performance. In fact, when a number of 3D jets are arrayed in a line, they resemble a 2D jet.

V. Summary and Conclusion. A two-dimensional, two-fluid model has been developed for characterization of the jet-contaminant interaction flow on a plane wall which occurs in chemical decontamination processes. Computer graphs are presented to show typical flow developments following the impingement process. The contaminant viscosity, the jet incidence angle, the jet velocity, and the jet size have been demonstrated to be important flow parameters in governing the flow pattern and the performance of the jet impingement for decontamination application.

The computer graphs show that in the high viscosity contaminant case a thick viscous layer develops quickly inside the flow region covered by the contaminant droplet and the jet stream is lifted off the impingement surface. From the computed results we have found that the optimum jet incidence angle for achieving the highest cleaning speed of the contaminant and the most efficient use of the jet fluid is in the range of 45° - 60°. An increase either in the jet velocity or in the jet size can greatly improve the cleaning speed. However, if the jet flow rate is the same, an increase in the jet velocity rather than the jet size will raise the efficiency of the use of jet fluid.

REFERENCES

1. G.I. Taylor, "Oblique Impact of a Jet on a Plane Surface," Phil. Trans. R. Soc. A, 260, 1966, pp. 96-100.

2. J.H. Michell, Phil. Trans. A, 181, 1890, pp. 389-431.

3. M.T. Scholtz and O. Trass, "Stagnation Flow-Velocity and Pressure Distribution," AIChE J., 16, No. 1, Jan 1979, pp. 82-96.

4. A. Rubell, "Computations of Jet Impingement on a Flat Surface," AIAA J., 18, No. 2, Feb 1980, pp. 168-175.

5. A. Rubell, "Computations of the Oblique Impingement of Round Jets Upon a Plan Wall," AIAA J., 19, No. 7, Jul 1981, pp. 863-871.

6. D.R. Kotansky and W.W. Bower," A Basic Study of the VTOL Ground Effect Problem for Planar Flow," J. Aircraft, 15, No. 4, Apr 1978, pp. 214-221.

7. J.N. Hunt, "Wave Formation in Explosive Welding," 1967.

8. J-M Vanden-Broeck, "Deformation of a Liquid Surface by a Impinging Gas Jet," SIAM J. Appl. Math., 41, No. 2, Oct 1981, pp. 306-309.

9. B.D. Nicholas, C.W. Hurt, and R.S. Hotchkiss, "SOLA-VOF: A Solution Algorithm for Transient Fluid Flow with Multiple Free Boundaries," Los Alamos Scientific Laboratory Report No. LA-8355, 1980.

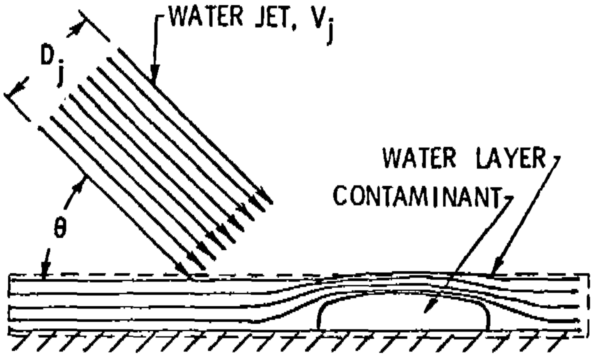


Figure 1. Flow configuration before impingement

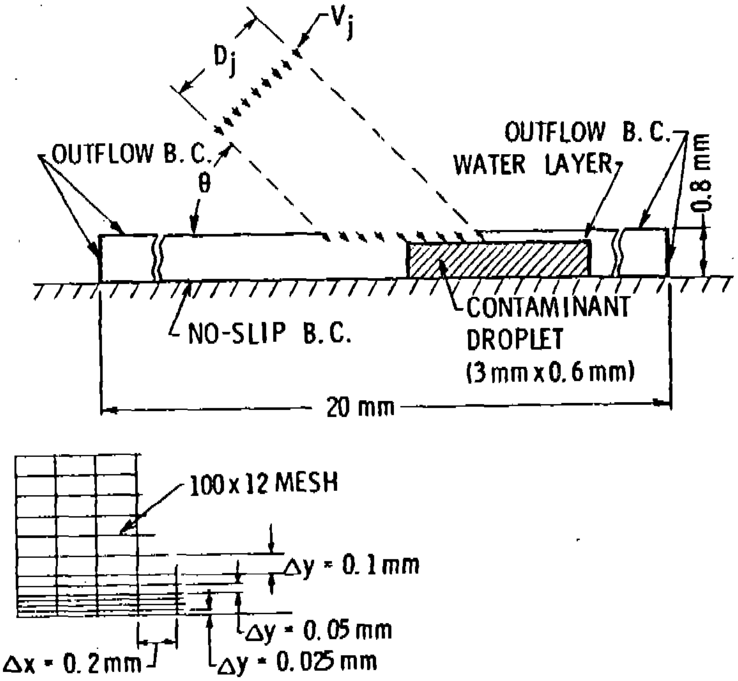


Figure 2. Model for numerical computation

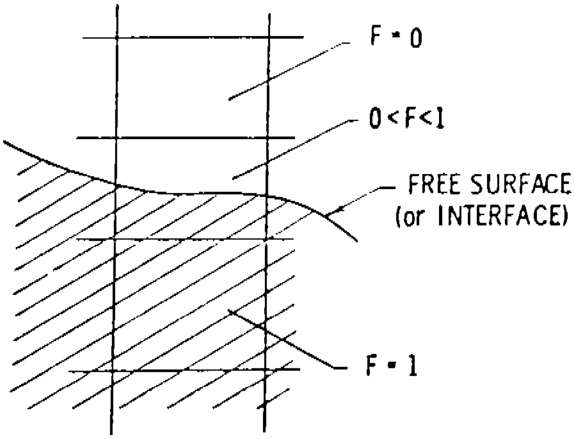


Figure 3. Interface between two fluids.

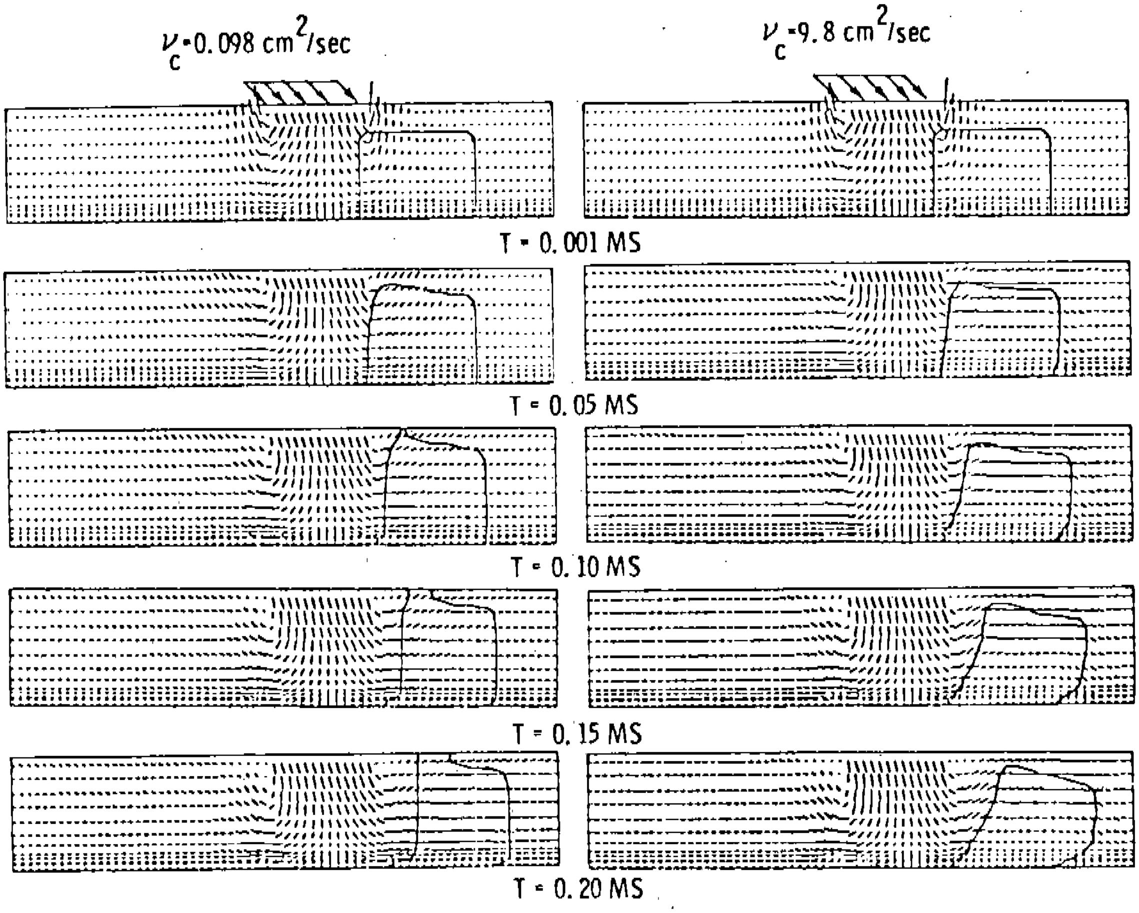


Figure 4. Flow patterns (two-fluid flow, ν_w = 0.0098 cm^2/sec, V_j = 5 m/sec, θ = 45^o)

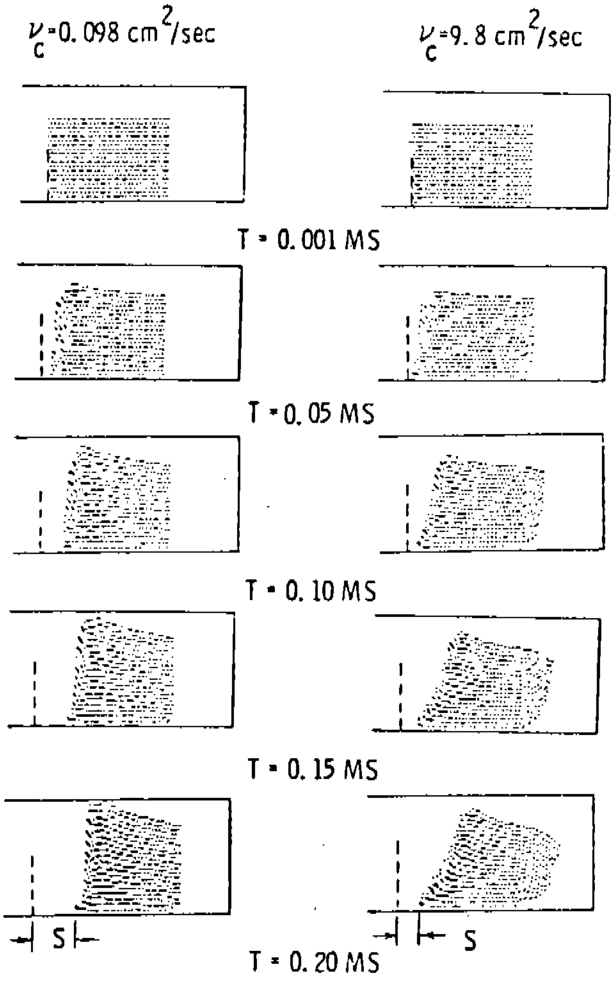


Figure 5. Evolution of contaminant droplets (V_j = 5 m/sec, θ = 45°)

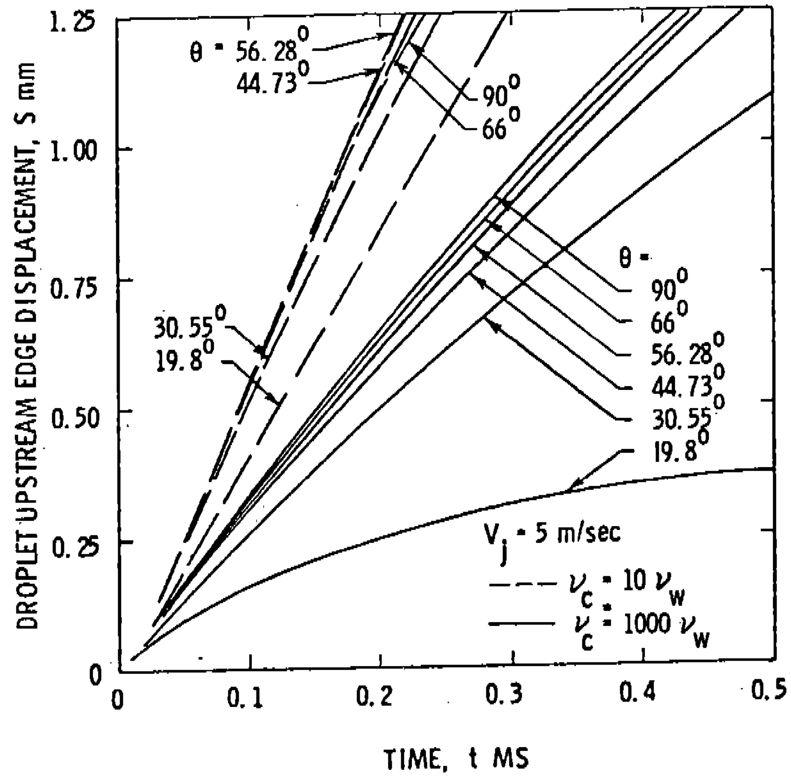


Figure 6. Droplet upstream edge displacement, S, vs. time after commencement of jet flow, t, for various jet incidence angles, θ (jet velocity V_j = 5 m/sec)

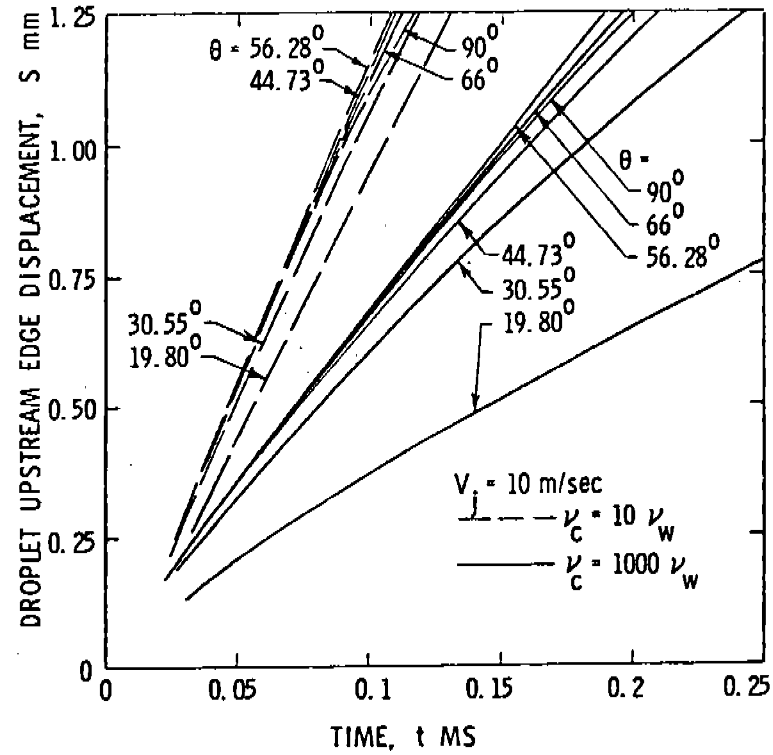


Figure 7. Droplet upstream edge displacement, S, vs. time after commencement of jet flow, t, for various jet incidence angles, θ (jet velocity V_j = 10 m/sec)

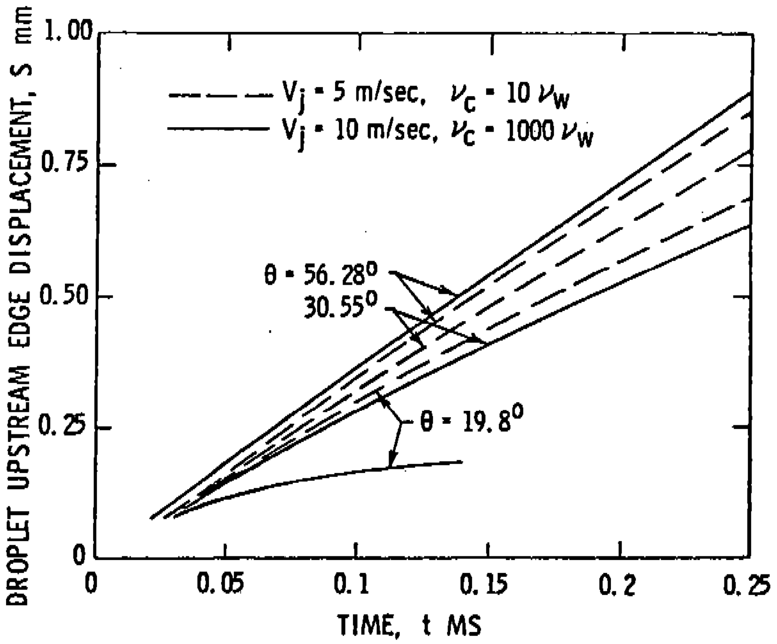


Figure 8. Droplet upstream edge displacement, S, vs. time after commencement of jet flow, t (water layer thickness is increased from 0.2 mm to 0.6 mm)

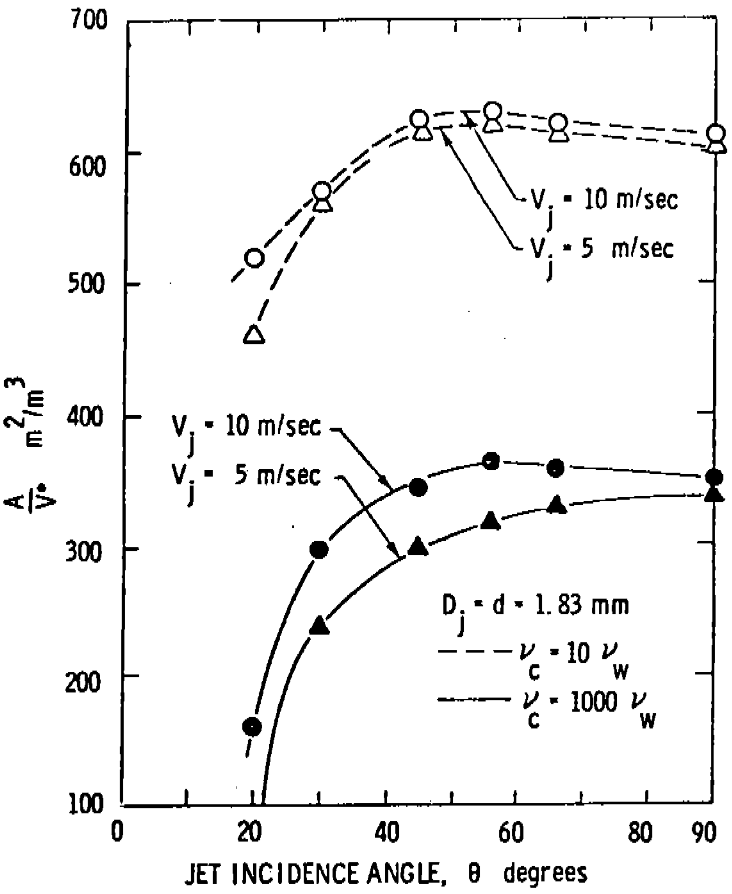
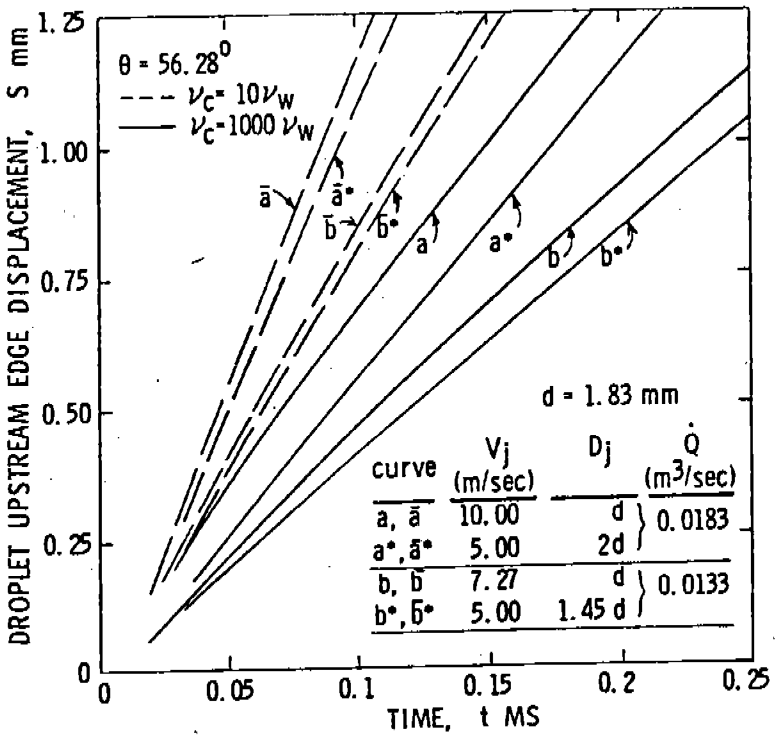


Figure 9. Nominal area cleaned per unit volume of jet fluid consumed, A/V*, for various jet incidence anyles, θ



curve	V_j (m/sec)	D_j	$\dot{Q}$ (m³/sec)
a, ā	10.00	d	0.0183
a*, ā*	5.00	2d	
b, b̄	7.27	d	0.0133
b*, b̄*	5.00	1.45 d	

Figure 10. Droplet upstream edge displacement, S, vs. time after commencement of jet flow, t, for various combinations of jet velocity and jet size

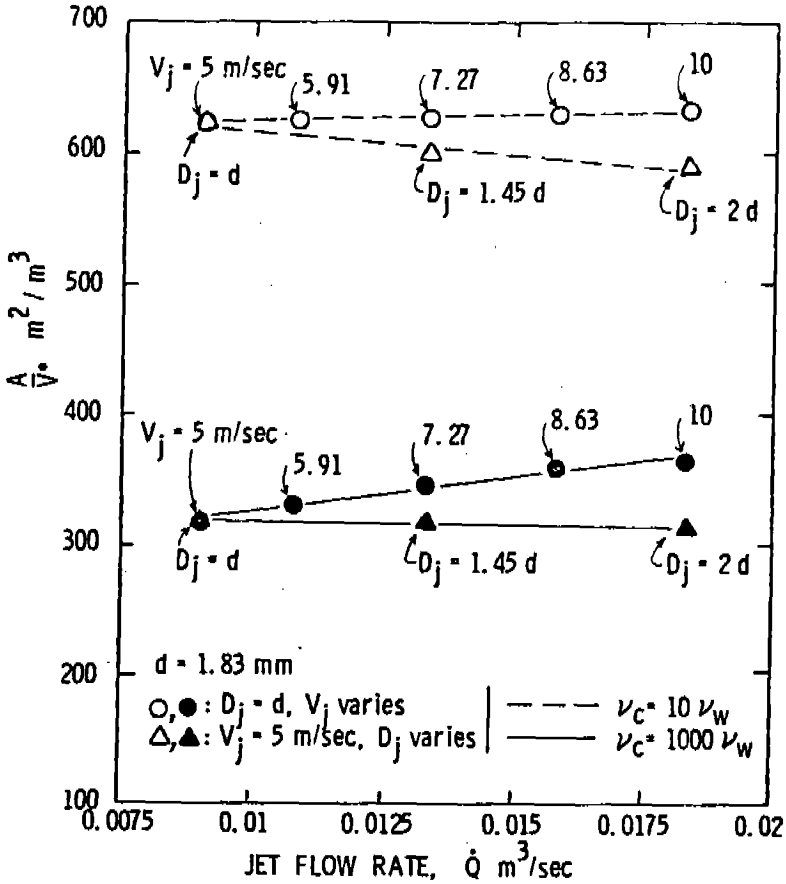


Figure 11. Nominal area cleaned per unit vlume of jet fluid consumed, A/V*, for various jet flow rate, $\dot{Q}$

NONUNIQUENESS AND STABILITY OF THE CONFIGURATION OF FLOW OF IMMISCIBLE FLUIDS WITH DIFFERENT VISCOSITIES

D. D. Joseph, K. Nguyen and G. S. Beavers
Department of Aerospace Engineering and Mechanics
University of Minnesota
Minneapolis, Minnesota 55455

March 1983

High viscosity liquids hate to work. Low viscosity liquids are the victims of the laziness of high viscosity liquids because they are easy to push around.

ABSTRACT. The arrangement of components in steady flow of immiscible liquids is typically nonunique. The problem of selection of arrangements is defined here and is studied by variational methods under the hypothesis that the realized arrangements are the ones which maximize the speed on exterior boundaries for prescribed boundary tractions, or the ones which minimize the tractions for prescribed speeds. The arrangements which minimize tractions also minimize the dissipation by putting low viscosity liquid in regions of high shear. The variational problem is used as a guide to intuition in the design and interpretation of experiments when results of analysis of stability are unavailable. In fact we always observe some kind of shielding of high viscosity liquid. This can occur by sheet coating in which low viscosity liquid encapsulates high viscosity liquid, or through the formation of rigidly rotating masses of high viscosity liquid which we call rollers. In other cases we get emulsions of low viscosity liquid in a high viscosity foam. The emulsions arise from a fingering instability. The low viscosity liquid fingers into the high viscosity liquid and then low viscosity bubbles are pinched off the fingers. The emulsions seem to have a very low effective viscosity and they shield the high viscosity liquid from shearing. In the problem of Taylor instability with two fluids low viscosity Taylor cells are separated by stable high viscosity rollers.

1. INTRODUCTION. We are interested in the flow of two immiscible liquids separated by an interface, driven by prescribed forces of the usual type. We call such motions bicomponent or two phase flows. In fact, we do not consider two phases of the same material, but of separate liquids which do not mix (oil and water, for example). To fix our problem, we specify the total volume of the two fluids and the individual volumes occupied by each one of them. Then our problem is to describe the motion and the spatial arrangement of each component. There is a high degree of nonuniqueness in such problems, even when the motion is steady and even when the region of flow is bounded and the Reynolds numbers for each of the fluids is very small. In some cases we may find a class of steady motions: no motions, Couette motions, Poiseuille motions, etc.,

which are uniquely determined by the data up to an arbitrary rearrangement of components (phases). The arrangement of components which is actually achieved in the flow is certainly connected to the problem of hydrodynamic stability. However, it has been suggested that in some problems the configurations which are achieved are those which in some sense extremalize the viscous dissipation. In fact a more precise statement of this apparently mystical idea can be formulated in terms of fluxes and forces. The idea is to maximize the flux for a given force (or to minimize the force for a given flux) over an admissible class of phase arrangements in a set of nonunique steady solutions. Our experiments show that something like this is going on. The arrangements of the components do in fact appear to be ones which extremalize in some mathematical sense. The extremalizing configurations are such as to minimize the shearing of high viscosity liquids by the spontaneous migration of low viscosity liquids into regions where the shearing is greatest.

The experiments reported in this paper seem to be of extraordinary interest in that they exhibit previously unknown types of fluid dynamics which may be typical for flows of immiscible liquids. Our experiments are visual and qualitative. In the future it would be good to systematically monitor torques and speeds and to undertake to correlate observations with systematic variations of geometric parameters.

2. EQUATIONS OF MOTION. The equations of motion in each liquid are the usual ones. We are going to study Navier-Stokes fluids in this paper but we are not yet certain that Navier-Stokes dynamics suffices to explain all that we have observed. The stress $\underline{T}$ is given by

$$(2.1) \qquad \underline{T} = -p\,\underline{1} + \underline{S}, \quad \underline{S} = 2\mu\,\underline{D}[\underline{u}]$$

where $\underline{D}$ is ½ of the symmetric part of the velocity gradient $\nabla\underline{u}$ and for all fluids in all regions of flow

$$(2.2) \qquad \text{div}\ \underline{u}_\ell = 0, \quad \ell=1,2\ ,$$

$$(2.3) \qquad \rho\frac{d}{dt}\underline{u}_\ell = -\nabla\Phi_\ell + \text{div}\ \underline{S}_\ell\ ,\ \underline{S}_\ell = 2\mu_\ell\ \underline{D}[\underline{u}_\ell]\ ,$$

$$(2.4) \qquad \Phi_\ell = \rho_\ell g\,Z + p_\ell$$

where g is gravity, Z increases against $\underline{g}$, ρ_1 and ρ_2 are the densities of the first and second fluids and μ_1 and μ_2 are the viscosities.

The interface between the regions 1 and 2 is called Σ and is given by

$$(2.5) \qquad f(\underline{x}(t),t) = 0$$

and, since this is an identity in t

$$(2.6) \qquad \frac{\partial f}{\partial t} + \underline{u} \cdot \nabla f = 0$$

where we have assumed that the normal component of velocity $d\underline{x}/dt$ of the surface Σ and the particles of fluid on either side of Σ are the same. In fact, the velocity $\underline{u}$ is continuous across Σ

$$(2.7) \qquad [\![\underline{u}]\!] = \underline{u}_1 - \underline{u}_2 = 0.$$

Now we state the conditions of continuity for the stresses across Σ. Let $\underline{n} = \nabla f/|\nabla f|$ be the normal to Σ and $\underline{\tau}_1$, $\underline{\tau}_2$ be two orthonormal vectors in Σ. The jump across Σ of the traction $[\underline{T}] \cdot \underline{n}$ satisfies

$$(2.8) \qquad [\![\underline{T}]\!] \cdot \underline{n} + \nabla_{II}\sigma + 2H\sigma\underline{n} = 0$$

where ∇_{II} is a surface gradient, 2H is the sum of the principal curvatures and σ is surface tension. Now, using (2.4) we introduce the head Φ,

$$(2.9) \qquad [\![\underline{T}]\!] = -[\![p]\!]\underline{1} + [\![\underline{S}]\!] = -[\![\Phi]\!]\underline{1} + g\,[\![\rho]\!]\,z_\Sigma\underline{1} + [\![\underline{S}]\!]$$

where z_Σ is the value of the coordinate z on Σ. We project (2.8), using (2.9), with $\underline{n}$ and $\underline{\tau}_1$, $\underline{\tau}_2$ to find that

$$(2.10) \qquad -[\![\Phi]\!] + \underline{n} \cdot [\![\underline{S}]\!] \cdot \underline{n} + g\,z_\Sigma[\![\rho]\!] + 2H\sigma = 0\,,$$

$$(2.11) \qquad \underline{\tau}_\ell \cdot [\![\underline{S}]\!] \cdot \underline{n} + \underline{\tau}_\ell \cdot \nabla_{II}\sigma = 0\,, \quad \ell=1,\,2\,.$$

This statement of conditions is completed by stating boundary and other auxilliary conditions. In general the position of the interface, f(x,t) = 0 is an unknown, to-be-determined quantity. In many cases, even after specifying the volumes occupied by each fluid, the solutions of the equations are not uniquely determined by the boundary data. In fact, there is a very high degree of non-uniqueness which we shall now specify more precisely.

3. HYDROSTATIC CONFIGURATIONS WHICH ARE NOT UNIQUE. Immiscible liquids of the same density will form spheres of one liquid in another. The steady distribution of such spheres, their sizes and their placement, seems not to be unique. In fact Plateau (1873) suspended masses of olive oil in a mixture of alcohol (lighter than oil) and water (heavier than oil) of the same density. By the latter device spheres of many centimeters were obtained. It is possible to have big spheres of oil in alcohol-water and big spheres of alcohol-water in oil. For such spheres we satisfy the equations of §2 with

$$\underline{u} = \underline{S} = \nabla_{II}\sigma = [\![\rho]\!] = 0$$

and

$$2H\sigma = [\![\Phi]\!] = [\![p]\!] = \frac{2\sigma}{R}$$

where R is the radius of each sphere. There is nothing here to determine the size and placement of spheres from given data.

The type of stationary configuration with bubbles of different sizes in immiscible fluids of matched density which was studied by Plateau is shown in Plate 3.1. This plate shows the result of matching the density of dibutyl phthalate with a glycerine (heavier) and water (lighter) solution.

The nonunique stationary configurations which are achieved by matching density appear to be stable to small disturbances. However, there is a tendency for bubbles which touch to collapse into one bubble. It may be true that there is a selection mechanism based on stability to large disturbances in which the stable configuration is the one which minimizes surface area. This type of criterion leads to large bubbles, even one large one, rather than many small ones.

4. NONUNIQUE PLANE COUETTE FLOW IN LAYERS. Another case of nonuniqueness of steady flow of two immiscible fluids is plane Couette flow in layers. Such flows are like heat conduction in layered materials and are characterized by alternate layers of fluids whose speed and shear stress is matched at each interface. The flowing fluid with velocity $\underline{u} = \underline{e}_x\, u_1(y)$, density ρ_1 and viscosity μ_1 occupies a height ℓ_1 per unit area in the plane of flow, the second fluid occupies a height ℓ_2 per unit area where $\ell_1 + \ell_2 = \ell$ is the total height of the layers and the two fluids may have any number of contiguous layers as in Fig. 4.1. The velocity functions

$$u(y) = a\,y + b \tag{4.1}$$

are linear in each layer and satisfy the conditions

$$\llbracket \mu \frac{du}{dy} \rrbracket = \mu_1 \frac{du_1}{dy} - \mu_2 \frac{du_2}{dy} = 0 ,$$

$$\llbracket u \rrbracket = u_1 - u_2 = 0, \tag{4.2}$$

$$u(0) = 0,$$

$$u(\ell) = U.$$

The normal stress balance is automatically satisfied by the fields (4.1), on flat interfaces, whether the surface tension is small or large. It matters not whether the densities are the same or different. If they are different then the steady layered flow exists, but is unstable, with all the heavy stuff falling to the bottom. If the densities are the same, no one knows what is the preferred arrangement of the layers. In fact Yih (1967) has shown that many of the configurations of flow with two layers are stable to small disturbances in the form of long waves and many are unstable to the same type of disturbance as the Reynolds number tends to zero. Stability depends on the viscosity ratio and the volume ratio. So non-uniqueness of layered Couette flow remains even after analysis of stability eliminates some steady arrangements. In the case of instability, Yih says that the interface becomes wavy. It would be interesting to study the bifurcation of layered Couette flow with flat interfaces into shear flow with wavy interfaces. If such solutions do bifurcate there would be a yet greater degree of nonuniqueness.

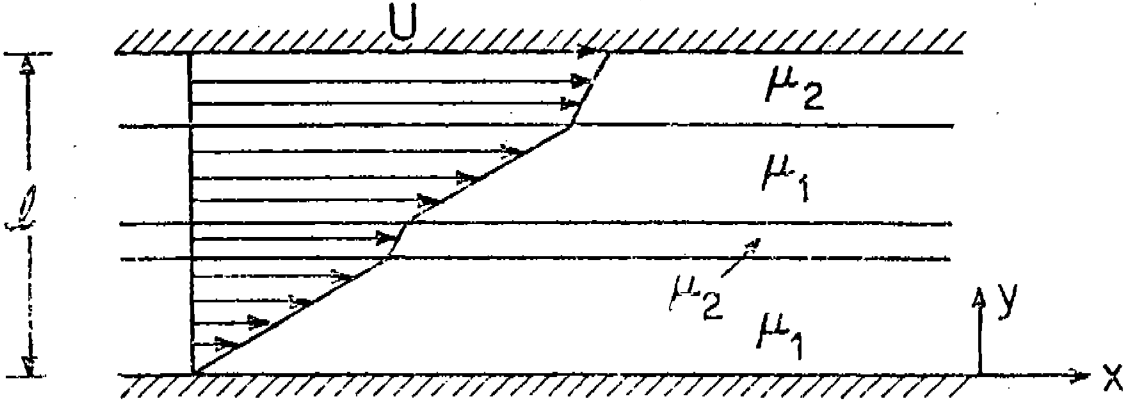


Fig. 4.1: Layered Couette flow. The high viscosity fluid is the one with viscosity $\mu_2 > \mu_1$. The two layers with fluid 2 have a total height equal to ℓ_2 and the two layers with fluid 1 have a total height of ℓ_1, $\ell_1 + \ell_2 = \ell$. The velocity profiles are linear functions of y, which increase in a continuous way from $u(0) = 0$ to $u(\ell) = U$.

5. NONUNIQUE PLANE POISEUILLE FLOW IN LAYERS. Another case of nonuniqueness of steady flow of two immiscible fluids is plane Poiseuille flow in layers. The set-up is the same as in §4, but u(y) is generated by a constant pressure gradient

$$\frac{dp}{dx} = -G$$

instead of by shearing from the top wall. Hence, in each layer

$$(5.1) \qquad \mu \frac{d^2u}{dy^2} = -G$$

where, at each interface $[\![u]\!] = [\![\mu \frac{du}{dy}]\!] = 0$ and $u(0) = u(\ell) = 0$. This problem has a continuum of solutions in which the fluid with viscosity μ_1 is in N layers of total height ℓ_1 separated by layers of fluid with viscosity μ_2 whose total height is ℓ_2. The heights of the constituent layers and their number are otherwise arbitrary. The remarks made about density differences in §4 apply in full force here.

Yih (1967) has also considered the stability of such flow of two fluids with different viscosities in two layers separated by one interface. He finds that Poiseuille flow in two layers is always unstable to long waves even for small Reynolds numbers tending to zero. Yih does not raise the problem of preferred arrangement of constituents. In fact it is probable that among the continuum of layered Poiseuille flows the one with the greatest stability is that for which the high viscosity fluid is centrally located (see Fig. 5.1). This configuration maximizes the mass flux for a given G, ℓ_1, ℓ_2 and requires the study of flows with at least three layers, not admitted into the analysis of Yih. The variational problem associated with the problem of the preferred arrangement of constituents is formulated in §11, and a solution of this problem for bicomponent Poiseuille flow in pipes is given in a paper by Joseph, Renardy and Renardy (1983) which is a companion to this one. We shall refer to this companion paper as JRR(1983).

6. NONUNIQUE POISEUILLE FLOW IN PIPES OF ARBITRARY CROSS-SECTION. Another case of nonuniqueness of steady flow of two immiscible fluids is Poiseuille flow in pipes of arbitrary cross-section. We shall imagine that the fluids have the same density but different viscosities. The cross-section of the pipe is called Ω and the coordinates in Ω are (y,z) with x increasing along generators of the pipes. The total cross-sectional area occupied by each of the two components Ω_1 and Ω_2, $\Omega_2 + \Omega_1 = \Omega$, and the pressure gradient (-G) are prescribed. We may find a continuum of phase arrangements satisfying

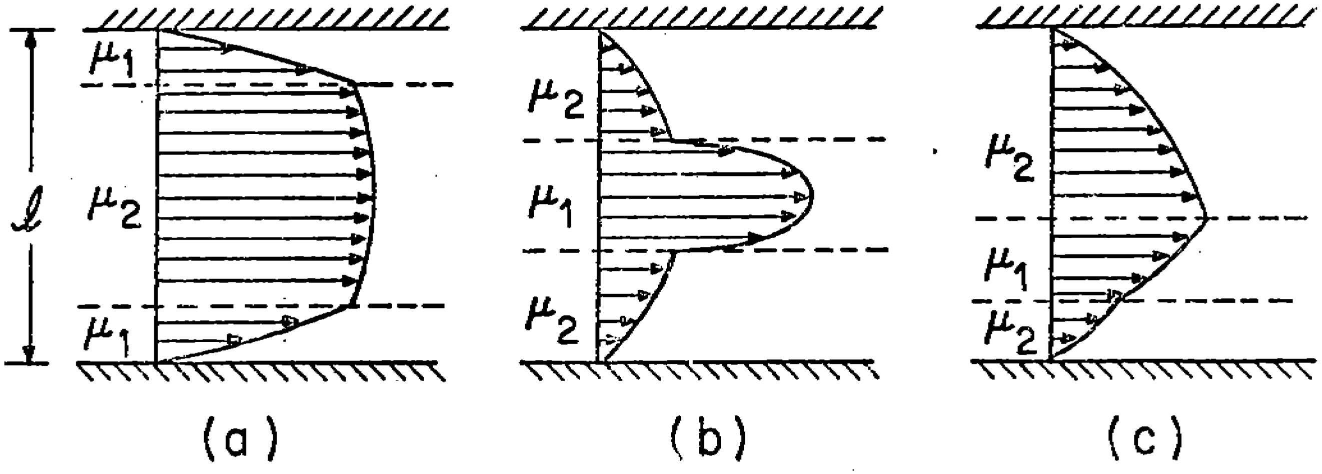


Fig. 5.1: Layered plane Poiseuille flow in three layers. The configurations in (a) and (b) are the centrally located ones. The configuration in (a) maximizes the mass flux for a given pressure gradient (-G) among the continuum of contiguous layers of high viscosity (μ_2) layers of total height ℓ_2 and low viscosity (μ_1) layers of total height ℓ_1; $\ell_1 + \ell_2 = \ell$.

$$\underline{u} = \underline{e}_x\, u(y, z), \qquad u(y,z)\Big|_{\partial\Omega} = 0,$$

$$\mu_\ell \left[\frac{\partial^2 u_\ell}{\partial_y^2} + \frac{\partial^2 u_\ell}{\partial_z^2}\right] = -G \;;\; \ell=1,\ 2\ .$$

At each interface we require that the velocity and the shear stress be continuous. The prescription of the shape and placement of the interfaces is left completely arbitrary, subject to the constraint of prescribed total area. The jump of the normal stress across interfaces vanishes automatically for solutions of Poiseuille type when the surface tension $\sigma = 0$. If $\sigma \neq 0$, the pressure will jump across each interface, and the value of the jump will be independent of (y,z), and balanced by a constant surface tension force; that is by constant interface curvatures. Hence, when $\sigma \neq 0$, the allowed interfaces are circles, or circular arcs terminating on boundaries. The number of such arcs and their placements are arbitrary.

There are experiments which suggest that in the flow of two fluids in a pipe there is a tendency for the low viscosity fluid to encapsulate the high viscosity one. The flows are such as to put the high viscosity component in the center of the pipe where the shears are smallest (see §10 and JRR(1983)).

7. NONUNIQUE COUETTE FLOW IN CIRCULAR RINGS BETWEEN ROTATING CYLINDERS. Another case of nonuniqueness of steady flow of two immiscible fluids is Couette flow in circular rings between rotating cylinders. We shall imagine that the fluids have the same density but different viscosities. We look for solutions in which the liquids are arranged in contiguous circular layers subject to a prescription of the total volume of each component liquid, (Fig. 7.1). Some solutions of this type will be given in §12. We look for solutions in each layer in the form

$$\underline{u} = \underline{e}_\theta\, v(r)$$

where the speeds of the inner (r=a) and outer (r=b) cylinders are prescribed

$$v(a) = \Omega_1 a,$$

$$v(b) = \Omega_2 b\ .$$

The v(r) that is required by the Navier-Stokes equations in each layer is of the form

$$v(r) = Ar + \frac{B}{r}\ .$$

The number and thickness of such circles are arbitrary, with the fixed volume constraint, and we can choose sets of {A} and {B} so that at each interface

$$[\![v]\!] = [\![\mu \frac{d(v/r)}{dr}]\!] = 0.$$

The normal stress equation may be satisfied by balancing pressure jumps against surface tension.

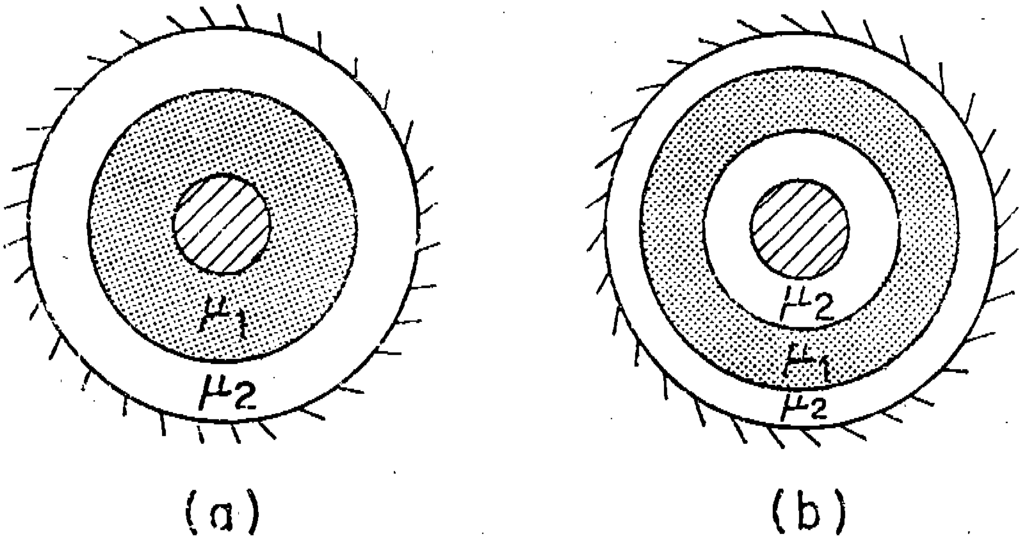


Fig. 7.1: Couette flow in circles of immiscible liquids of the same density but different viscosity: (a) two layers, (b) three layers with the same total volume of μ_1 and μ_2. The number and thickness of these layers are arbitrary.

8. MORE GENERAL EXAMPLES OF NONUNIQUENESS IN THE STEADY FLOW OF IMMISCIBLE LIQUIDS. Nonunique arrangements of components in bicomponent flows appear to be a general property going far beyond the simple examples exhibited in the previous sections.

For example we could generate nonunique two dimensional "Poiseuille" flows by, say, a wavy perturbation of the solid boundary. We should then of course be obliged to show how the hydrodynamics and interfaces perturb under the boundary perturbation.

As another example we may consider how the circular flows in Fig. 7.1 perturb when there is a small density difference $[\![\rho]\!] = \varepsilon \neq 0$. We expect to see slightly displaced circles.

As another example consider the problem of one fluid displacing another in a pipe (see Fig. 8.1). This problem is a model for studies of the motion of the contact line. If there are such solutions, they are not unique because there are also at least layered Poiseuille flows. If $\mu_1 < \mu_2$ the less viscous fluid will finger into the more viscous fluid. In this case, and also when $\mu_2 < \mu_1$, the configuration shown in Fig. 8.1 is probably unstable.

Another example, due to F. Busse (1982), is described by him as follows: "Convection in a fluid layer heated from below with two immiscible components A and B of the same average density exhibits different states of motion. Besides a solution describing convection in sublayers other solutions in which fluid B is surrounded by streamlines of fluid A or vice versa are possible."

In our experiments, we see many different configurations. There are persistent solutions with bubbles of low viscosity fluid in high viscosity foams, which are steady in some average sense.

We conclude that, unlike single component flow, there is a pervasive lack of uniqueness in the flow of immiscible liquids even when the Reynolds number is small and even when the region of flow is bounded.

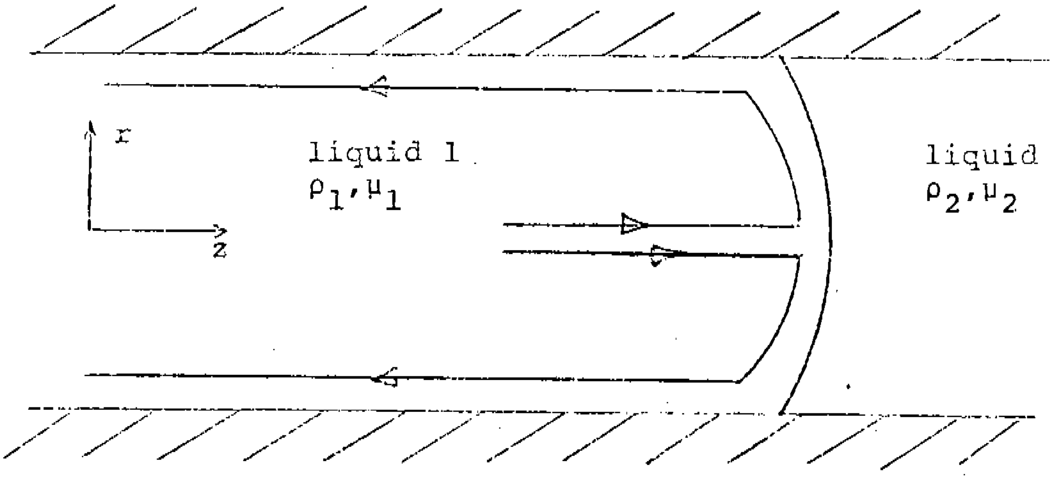


Fig. 8.1: One fluid displaces another in a horizontal pipe. This configuration is probably unstable.

9. <u>PHYSICAL MECHANISMS FOR SELECTING THE ARRANGEMENT OF COMPONENTS PREFERRED IN THE FLOW OF IMMISCIBLE LIQUIDS</u>. The natural thing to do when faced with nonuniqueness is to study *stability*. It is probable that most of the possible steady configurations of flow are unstable. Our experiments show that there is a definite tendency for flows of immiscible liquids to arrange themselves in such a way as to shield the high viscosity components from intense shearing. We shall call this tendency for low viscosity liquids to migrate into the regions of high shears an *encapsulation* instability. A *fingering* instability which is responsible for the formation of emulsions is also important in the context of encapsulation. These types of instability are important because they represent a type of self-lubrication principle.

There is an irresistible temptation to invoke variational principles for selection criteria for the arrangements of components which are preferred in bicomponent flow. These principles are suggested by various results associated with extremalizing the dissipation. Such principles, as a strict statement of the underlying physics, are oversimplified. In fact, exact variational principles are ambiguous. One can state different principles leading to different selection rules. We are going to postulate some such "principles", because they are mathematically interesting, very suggestive of true underlying physics and provide the right ambience for the discussion of this problem. So in this case as in many others we do not resist irresistible temptation. However, in JRR(1983) a good variational problem is stated, solved, shown to agree with experiments (the experimental results that we know; there are surely others) and to be in partial agreement and partial contradiction with the analysis of stability.

The big interest in variational ideas for us was that it guided our intuition in the design of experiments in situations in which analysis is hard, impossible or in any event not available. Probably the coarse statement that the hydrodynamics shields the high viscosity component from shearing is more nearly correct than any exact variational embodiment of it.

10. ENCAPSULATION INSTABILITIES AND EXTREMAL PRINCIPLES FOR PIPE FLOW. In pipe flow of two liquids with different viscosities under an applied pressure drop, the low viscosity liquid will tend to encapsulate the high viscosity liquid. If the effects of gravity are negligible, the phases will arrange themselves so whatever may have been the initial configuration, the high viscosity phase will ultimately be centrally located (as in Fig. 5.1(a)). This property has been convincingly demonstrated in experiments with very viscous viscoelastic liquids (polymer melts) by Southern and Ballman (1973), Everage (1973), Lee and White (1974), Williams (1975), and Minagawa and White (1975), as well as in the flow of oil and water, in which the water migrates to the pipe wall, forming a lubrication layer, studied by Charles and Redberger (1962), Hasson, Mann and Nir (1970), and Yu and Sparrow (1967).

In the experiments of Southern and Ballman (1973) encapsulation of the type exhibited in Fig. 10.1 is documented. Everage (1973) shows a photograph of complete encapsulation with a centrally located high viscosity nylon completely encapsulated by an annular ring of low viscosity fluid.

Theoretical explanations of the slow envelopment phenomenon have up to now been based upon extremalizing energy dissipation as originally suggested by Southern and Ballman (1973). Maclean (1973), who considered planar layered flow, and Everage (1973), who studied a cylindrical geometry, both invoked a variational principle to show that the phase configuration with the high-viscosity component centrally located is favored over several other configurations.

We are going to state a general variational criterion which presumably reduces to the ones first discussed for special cases. We say that realized flows arrange the two components so as to maximize the speeds when the tractions on exterior boundaries are fixed, or to minimize the tractions when the speeds are prescribed. For pipe flow we arrange the two flowing components to maximize the mass flux when the pressure gradient is prescribed or to minimize the pressure gradient when the mass flux is prescribed. The extremalizing arrangement is the one with high viscosity liquid in the center (JRR(1983)). We could draw an analogy with turbulent flow in which it is shown that among all the turbulent solutions, the laminar one maximizes the mass flux for a given pressure drop. In fact the extremalizing solution is not always stable. The results (JRR(1983)) depend on the volume ratio with stable flows characterized by narrow layers of less viscous fluid on the outside.

If we suppose that nature's design corresponds to man's desire, we could hope to realize the solution of the problem of maximizing the flux of oil down a pipe of fixed radius by lubricating the pipe wall with water. In fact the solution of the design problem is a stable one according to the calculations give in JRR(1983). The design problem is as follows:

Water and oil flow steadily along a horizontal pipe of radius R. The lower viscosity (μ_2) water completely encapsulates the higher viscosity (μ_1) oil, which flows as a central core of radius a.

The flow is described by the usual equation, with

$$u_2 = 0 \text{ at } r = R$$

and

$$\left.\begin{aligned} u_2 &= u_1 \\ \mu_2 \frac{du_2}{dr} &= \mu_1 \frac{du_1}{dr} \end{aligned}\right\} \quad \text{at } r = a.$$

The velocity distributions in the two fluids are:

$$u_1 = \frac{R^2}{4\mu_1}(G)\left(\frac{a^2}{R^2}\right)\left\{\frac{\mu_1}{\mu_2}\left(\frac{R^2}{a^2} - 1\right) + \left(1 - \frac{r^2}{a^2}\right)\right\}, \quad 0 \leq r \leq a$$

and

$$u_2 = \frac{R^2}{4\mu_1}(G)\left(\frac{a^2}{R^2}\right)\left(\frac{\mu_1}{\mu_2}\right)\left(\frac{R^2}{a^2} - \frac{r^2}{a^2}\right), \quad a \leq r \leq R,$$

where

$$G \equiv (-\frac{dp}{dx}).$$

The volume flow rate of oil (Q_1) is then

$$\frac{Q_1}{\tilde{Q}} = \frac{a^4}{R^4}\left\{1 + 2(\frac{\mu_1}{\mu_2})(\frac{R^2}{a^2} - 1)\right\}$$

where $\tilde{Q} = \frac{R^4}{8\mu_1}(G)$ is the volume flow rate of the single phase (μ_1) when it fills the entire pipe. For fixed R and G, the volume flow rate Q_1 has a maximum when

$$\frac{a}{R} = \left[\frac{\mu_1/\mu_2}{2(\frac{\mu_1}{\mu_2}) - 1}\right]^{\frac{1}{2}},$$

and the maximum value is then

$$\left.\frac{Q_1}{\tilde{Q}}\right)_{max} = \frac{(\frac{\mu_1}{\mu_2})^2}{2(\frac{\mu_1}{\mu_2}) - 1}.$$

For large values of the viscosity ratio μ_1/μ_2, as would occur for example in an oil/water configuration, the maximum volume flow rate Q_1 is obtained when $a \approx \frac{R}{\sqrt{2}}$, and is given by

$$\left.\frac{Q_1}{\tilde{Q}}\right)_{max} \approx \frac{1}{2}(\frac{\mu_1}{\mu_2}).$$

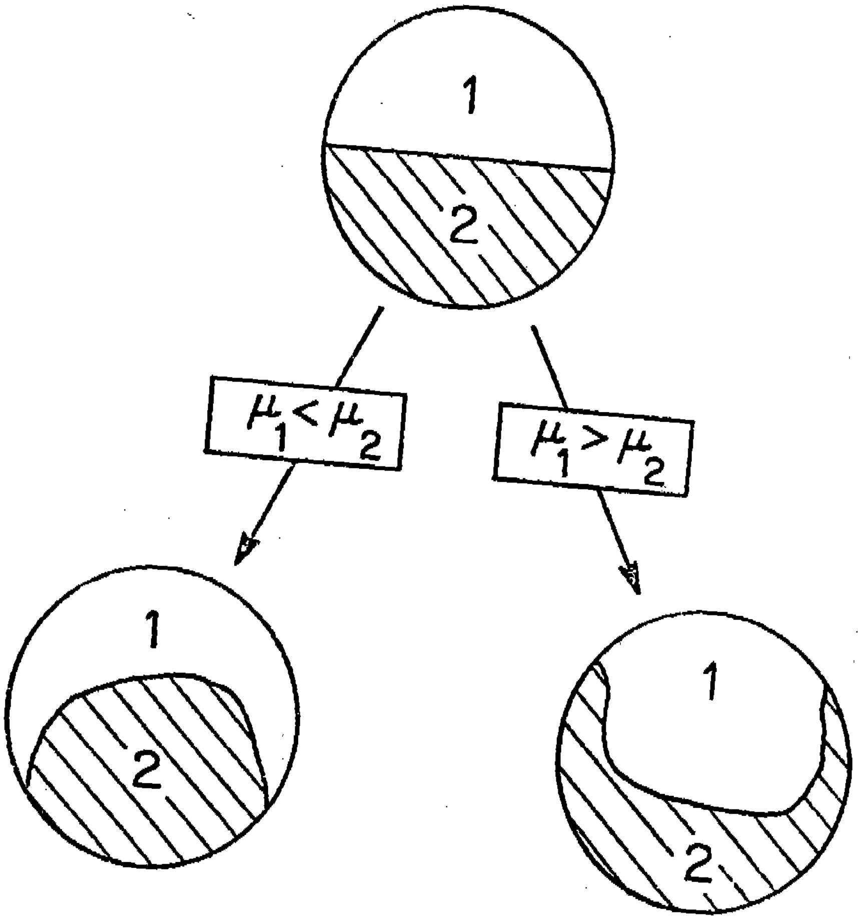


Fig. 10.1: Encapsulation of high viscosity liquid by low viscosity liquid in pipe flow of polymer melts (after Southern and Ballman, 1973).

The volume flow rate of the lower viscosity fluid is

$$\frac{Q_2}{\tilde{Q}} = \left(\frac{\mu_1}{\mu_2}\right)\left(1 - \frac{a^2}{R^2}\right)^2 .$$

Thus the volume flow rate Q_2 required to maximize Q_1 is

$$\left.\frac{Q_2}{\tilde{Q}}\right)_{\max Q_1} = \left(\frac{\mu_1}{\mu_2}\right)\left[\frac{\frac{\mu_1}{\mu_2} - 1}{2\left(\frac{\mu_1}{\mu_2}\right) - 1}\right]^2 .$$

For large values of $\frac{\mu_1}{\mu_2}$ this reduces to

$$\left.\frac{Q_2}{\tilde{Q}}\right)_{\max Q_1} \approx 1/4\left(\frac{\mu_1}{\mu_2}\right)$$

or

$$\left.\frac{Q_2}{\tilde{Q}}\right)_{\max Q_1} \approx \tfrac{1}{2}\left(\frac{Q_1}{\tilde{Q}}\right)_{\max} .$$

11. EVOLUTION OF THE ENERGY AND BALANCE OF POWER OF THE EXTERIOR TRACTIONS IN THE BICOMPONENT FLOW OF IMMISCIBLE LIQUIDS. The following energy equation due to E. Dussan V governs the flow of immiscible incompressible fluids (see Joseph, 1976, for a full discussion):

$$\frac{d}{dt}[\mathcal{E} + \mathcal{P} + \int_{\Sigma} \sigma d\Sigma] = \int_{\Sigma} \frac{d\sigma}{dt}\, d\Sigma$$

$$+ \oint_{\partial\Sigma} \sigma \underline{t}\cdot\underline{U} d\ell + \int_{\partial\mathcal{V}} \underline{u}\cdot(\underline{T}\cdot\underline{n}) \tag{11.1}$$

$$+ \int_{\mathcal{V}} \mathrm{tr}\ \underline{T}\ \underline{D}[\underline{u}] .$$

The terms of (11.1) need to be defined and discussed. The definitions of volumes, surfaces, distinguished directions and velocities are defined under Fig. 11.1. The stress $\underline{T}$ is given by (2.1), $\underline{D}[\underline{u}]$ is the stretching tensor for $\underline{u}$,

$\mathcal{E} = \int_{\mathcal{V}} \tfrac{1}{2}\rho|\underline{u}|^2$ (kinetic energy),

$\mathcal{P} = \int_{\mathcal{V}} \rho g z$ (potential energy),

$\oint_{\Sigma} \sigma\, d\Sigma$ (surface energy),

$\frac{d\sigma}{dt}$ (derivative of surface tension following a particle in the surface)

$\oint_{\partial\Sigma} \sigma\underline{t}\cdot\underline{U}d\ell$ (power or working of the contact line)

$\int_{\partial\mathcal{V}} \underline{u}\cdot(\underline{T}\cdot\underline{n})$ (power or working of the traction vector $\underline{T}\cdot\underline{n}$ on the exterior boundary)

$\int_{\mathcal{V}} \mathrm{tr}\ \underline{T}\ \underline{D}[\underline{u}]$ (stress power or dissipation)

For incompressible Newtonian fluids the dissipation is in the form

$$\int_{\mathcal{V}} 2\mu\ \mathrm{Tr}\ \underline{D}^2[\underline{u}] .$$

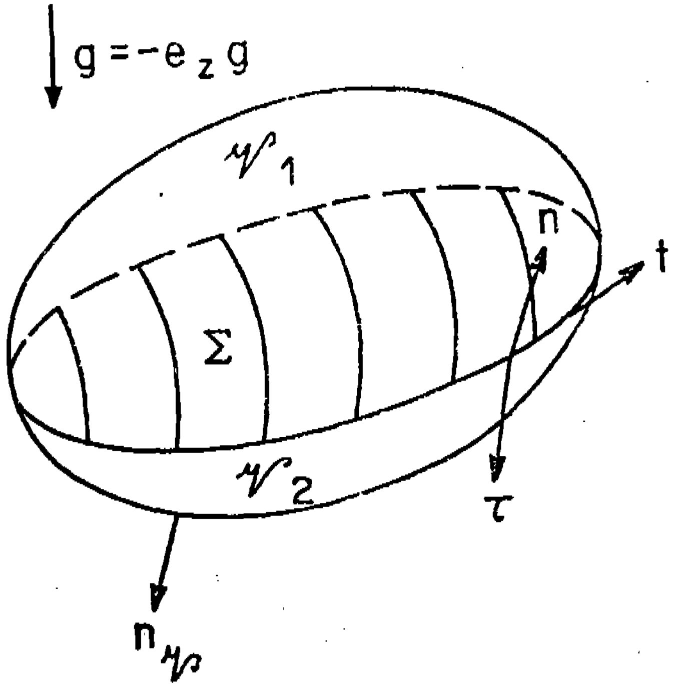


Fig. 11.1: Volumes, surfaces and distinguished directions for two fluids separated by an interface.

$\mathscr{V}_1$ and $\mathscr{V}_2$ are the volumes occupied by the two fluids,

$\mathscr{V} = \mathscr{V}_1 \cup \mathscr{V}_2$,

$\underline{n}_i$ is the outward normal to $\partial\mathscr{V}_i$ $(i = 1, 2)$,

$\Sigma = \partial\mathscr{V}_1 \cap \partial\mathscr{V}_2$ is the interface separating the two fluids,

$\underline{n}$ is the normal to Σ pointing from $\mathscr{V}_2$ to $\mathscr{V}_1$,

$\partial\mathscr{V} = \partial\mathscr{V}_1 \cup \partial\mathscr{V}_2 - \Sigma$ is the exterior boundary; parts of this boundary may be made of rigid solids,

$\underline{n}$ is the outward normal on $\partial\mathscr{V}$,

$\partial\Sigma = \Sigma \cap \partial\mathscr{V}$ is the contact line; ℓ is an arc length on this line,

$\underline{t}$ is the tangent vector on $\partial\Sigma$,

$\underline{\tau} = \underline{t} \wedge \underline{n}$ is the normal to $\partial\Sigma$ on Σ,

$\underline{u}_i$ is the velocity of the fluid in $\mathscr{V}_i$ $(i = 1, 2)$,

$\underline{U}$ is the velocity of a point of the contact line.

The steady flow of immiscible fluids satisfies (11.1) with time derivatives set to zero. If $\sigma = 0$, $\underline{U} = 0$ or if there is no contact line as in some flows of two fluids in pipes or between cylinders, then for Newtonian fluids

$$(11.2) \qquad \int_{\partial\mathcal{V}} \underline{u}\cdot(\underline{T}\cdot\underline{n}) = \int_{\mathcal{V}} 2\mu\underline{D}[\underline{u}] : \underline{D}[\underline{u}]$$

One selection rule which may be postulated is as follows. The realized placements of $\mathcal{V}_1$ and $\mathcal{V}_2$ within $\mathcal{V}$ is the one which maximizes the speeds u for prescribed tractions on the exterior boundary $\partial\mathcal{V}$ or minimizes tractions for prescribed speeds. Joseph, Renardy and Renardy (1983) have shown that this is a well-defined problem with a definite solution in some cases. The selection rule could be stated in terms of fluxes and forces instead of speeds and tractions. The selection rule requires that we first specify a class of nonunique flows, say layered Poiseuille flows, before we seek the optimum placements of components. Maximizing speeds is the same as maximizing dissipation for prescribed tractions. Minimizing tractions for given speeds requires that we minimize the dissipation. In Poiseuille flows it is natural to fix the pressure gradient (tractions) and maximize the flux (speeds). In contrast, in Couette flows we specify the speeds (angular velocity) and minimize the shear stresses (torques) at the boundary.

12. VARIATIONAL PROBLEMS FOR COUETTE FLOW IN PLANE AND CIRCULAR LAYERS AND FOR LAYERED POISEUILLE FLOW. Consider the set of layered Poiseuille flows discussed in §5. In Fig. 5.1 we exhibited some examples of three layer configurations. In general, the layered Poiseuille flows are uniquely determined up to arrangement. This means the total number N of layers and the size of the layers, subject to total height constraints, are left undetermined. Now we shall write the energy balance (11.2) for layered plane Poiseuille flow. The volume here is a plane area of channel height ℓ and length L (along the axis x). Since u vanishes on the solid walls the integral on the left of (11.2) is over planes perpendicular to x at x and x + L.

$$\int_{\partial\mathcal{V}} \underline{u}\cdot(\underline{T}\cdot\underline{n}) = \int_0^{\ell} [-uT^{<xx>}\Big|_x + uT^{<xx>}\Big|_{x+L}]dy$$

$$= (-p_{L+x} + p_x)\int_0^{\ell} udy = -\frac{\Delta p}{L}LQ = GLQ$$

where

$$Q = \int_0^{\ell} u dy \quad \text{is the volume flux.}$$

On the other hand

$$2\int \mu \underline{D}:\underline{D} = \tfrac{1}{2}\int \mu u'^2(y)\, d\mathcal{V}$$

$$= \frac{L}{2}\int_0^{\ell} \mu u'^2(y)\, dy.$$

Hence

$$(12.1) \qquad 2GQ = \int_0^{\ell} u u'^2(y)\, dy.$$

We next recall that between 0 and ℓ are N contiguous layers with fluids of different viscosities. We can suppose that the layer nearest the bottom is occupied by a fluid of viscosity μ_1, the next layer has μ_2, then μ_1 again, and so on. So besides the total number N of layers and their sizes, we need to know if μ_1 is the larger viscosity. We suppose that the total volume (height) of high viscosoty (μ_+) fluid is given as ℓ_+ and ℓ_- is the volume of low viscosity fluid and $\ell_+ + \ell_- = \ell$. The N layers are divided into intervals

$$[0,y_{(1)}],[y_{(1)},y_{(2)}],[y_{(2)},y_{(3)}], \ldots [y_{(N-1)},\ell].$$

With G given we maximize

$$(12.2) \qquad Q = \int_0^{Y(1)} u_{(1)}(y)\,dy + \int_{Y(1)}^{Y(2)} u_{(2)}(y)\,dy + \ldots + \int_{Y(N-1)}^{\ell} u_{(N)}(y)\,dy$$

where

$$(12.3) \qquad 2QG = \mu_1 \int_0^{Y(1)} u'^2_{(1)}(y)\,dy + \mu_2 \int_{Y(1)}^{Y(2)} u'^2_{(2)}(y)\,dy$$

$$+ \mu_1 \int_{Y(2)}^{Y(3)} u'^2_{(3)}(y)\,dy \ldots .$$

Now we change variables:

$$(12.4) \qquad \begin{aligned} u_{(1)} &= -\frac{G}{\mu_1} v_{(1)}, \\ u_{(2)} &= -\frac{G}{\mu_2} v_{(2)}, \\ u_{(3)} &= -\frac{G}{\mu_1} v_{(3)} \end{aligned}$$

where

$$(12.5) \qquad v'_{(n)} = -y + Y, \quad y \in [0,\ell]$$

satisfies (5.1), $[\![\mu u']\!] = 0$ at each interface and there is one $Y \in (0,\ell)$ such that $v'_{(n)}(Y) = 0$. Inserting this change of variables into (12.3) we get:

$$(12.6) \qquad \frac{2Q}{G} = \frac{1}{\mu_1}\int_0^{Y(1)} (y - Y)^2 dy + \frac{1}{\mu_2}\int_{Y(1)}^{Y(2)} (y - Y)^2 dy$$

$$+ \frac{1}{\mu_1}\int_{Y(2)}^{Y(3)} (y - Y)^2 dy + \ldots$$

We maximize this by arranging the layers so that the high viscosity in the denominator is associated with the smallest value of $(y - Y)^2$. This is clearly the arrangement of (a) of Fig. (5.1), so $N = 3$, and μ_1 is the low viscosity μ_-.

The same considerations, but in more complicated form, enter into the rigorous solution of this problem for pipe flow given by JRR(1983).

Couette flow differs from Poiseuille flow in that it is perhaps more natural to prescribe the speed of exterior boundaries and to minimize the torque. We first note that for the Couette flow between cylinders in N rings and layers which is shown with N = 2 and 3 in Fig. 7.1 we have

$$\underline{e}_x \cdot (\underline{T}\cdot\underline{n}) = T^{<\theta x>} = 0$$

so that

$$(12.7) \qquad \int_{\partial\mathcal{V}} \underline{u} \cdot (\underline{T}\cdot\underline{n}) = \int_{\partial\mathcal{V}} v(r)\underline{e}_\theta \cdot (\underline{T}\cdot\underline{n})$$

$$= 2\pi \{b\Omega_2 T^{<r\theta>}_{(b)} - a\Omega_1 T^{<r\theta>}_{(a)}\} = (\Omega_2 - \Omega_1)M$$

where Ω_2 and Ω_1 are the angular speeds of the outer and inner cylinders, and $M = 2\pi r T^{<r\theta>}(r)$ is the torque and it is constant for $a \le r \le b$. It is clear from (11.2) that $(\Omega_2 - \Omega_1)M$ is positive and

$$
\begin{aligned}
M(\Omega_2 - \Omega_1) &= \frac{\mu}{2}\int_a^b r^3\left[\frac{d(v/r)}{dr}\right]^2 dr \\
&= \frac{\mu_1}{2}\int_a^{r(1)} r^3\left[\left(\frac{v}{r}\right)'\right]^2 dr + \frac{\mu_2}{2}\int_{r_{(1)}}^{r(2)} r^3\left[\left(\frac{v}{r}\right)'\right]^2 dr \\
&+ \frac{\mu_1}{2}\int_{r_{(2)}}^{r(3)} r^3\left[\left(\frac{v}{r}\right)'\right]^2 dr + \dots
\end{aligned}
\tag{12.8}
$$

and

$$\underline{u} = \underline{e}_\theta v(r), \quad [\![\rho]\!] = 0 \quad \text{and}$$

$$[\![\mu(\frac{v}{r})']\!] = [\![v]\!] = 0 \quad \text{at } r_{(1)},\ r_{(2)}, \dots .$$

We want to choose the arrangements of layers and the placement of viscosities so as to minimize $|M|$ when Ω_2 and Ω_1 are prescribed. It will suffice to solve this problem for the two layer configuration shown in Fig. 7.1. We may always consider the problem posed for two adjacent layers. We find that the minimizing solution has the lower viscosity fluid on the inside. We conclude that N = 2 with more viscous fluid on the outside and less viscous fluid inside.

Suppose $r_{(1)} = d$ is the interface between two layers at $r = a$ and $r = b$. The fluid with viscosity μ_2 occupies the region $a \leq r \leq d$, the fluid with viscosity μ_1 occupies the region $d \leq r \leq b$. We find that in $d \leq r \leq b$

$$v_1 = A_1 r + B_1/r \ ,$$

$$A_1 = \{\mu_2(\frac{\Omega_1}{b^2} - \frac{\Omega_2}{d^2}) + \mu_1\ \Omega_2(\frac{1}{d^2} - \frac{1}{a^2})\}/q \ ,$$

$$q \overset{\text{def}}{=} \mu_2(\frac{1}{b^2} - \frac{1}{d^2}) + \mu_1(\frac{1}{d^2} - \frac{1}{a^2}) \ ,$$

$$B_1 = (\Omega_2 - \Omega_1)\mu_2/q,$$

and in $a \le r \le d$:

$$v_2 = A_2 r + B_2/r,$$

$$A_2 = \{\mu_2\Omega_1(\frac{1}{b^2} - \frac{1}{d^2}) + \mu_1(\frac{\Omega_1}{d^2} - \frac{\Omega_2}{a^2})\}/q,$$

$$B_2 = \mu_1 B_1/\mu_2.$$

Since

$$M(\Omega_2 - \Omega_1) = \frac{\mu_2}{2}\int_a^d r^3\left[(\frac{v_2}{r})'\right]^2 dr + \frac{\mu_1}{2}\int_d^b r^3\left[(\frac{v_1}{r})'\right]^2 dr$$

(12.9)

$$= \frac{a^2b^2(\Omega_2 - \Omega_1)^2(ka^2 + b^2)\mu_1\mu_2}{(b^2 - a^2)(b^2\mu_1 + ka^2\mu_2)} \ ,$$

where $k = (b^2 - d^2)/(d^2 - a^2)$, is positive, we may without losing generality, consider the case for which M and $\Omega_2 - \Omega_1$ are positive. It is immediate that $M(k,\mu_1,\mu_2)$ is a monotonically increasing function of μ_2, from zero at $\mu_2 = 0$ to

$$b^2(\Omega_2 - \Omega_1)^2(ka^2 + b^2)\mu_1/(b^2 - a^2)k.$$

It follows from monotonicity that M is larger when $\mu_2 > \mu_1$ than when $\mu_2 < \mu_1$. So we minimize the torque by putting the lower viscosity fluid $\mu_- = \mu_2$ on the inner cylinder at radius $r = d$.

The situation when one of the liquids occupies an infinite region has to be treated separately. For example, when $b \to \infty$ and $\Omega_2 \to 0$ then

$$(12.10) \quad |M| = \frac{|\Omega_1| a^2 d^2 \mu_1 \mu_2}{(\mu_2 - \mu_1)a^2 + \mu_1 d^2},$$

which for a fixed d is smaller when $\mu_2 > \mu_1$.

So in every pair of layers the arrangement which minimizes the torque has the low viscosity liquid on the inside. It follows that optimal arrangements of layers for minimum torque is the one with two layers and the less viscous fluid on the inner cylinder.

The preferred arrangement for the problem of plane Couette flow in which the velocity is U at $y = \ell$ and zero at $y = 0$ may be studied using (11.2) in reduced form

$$(12.11) \quad UT^{<xy>} = \int_0^\ell u'^2(y)\,dy.$$

It is not hard to verify that the value of the integral on the right of (12.11) is independent of the number and size of layers if the total volume ℓ_+ and ℓ_- of the high and low viscosity fluids is prescribed. It follows that the variational problem for the preferred arrangement of layers in plane Couette flow has no solution.

13. STABILITY. Ultimately the only satisfactory theoretical approach to the problem of selection is through stability. There are a few papers which treat the problem of stability of immiscible liquids with different viscosities. Of these, the papers of Yih (1967) and Joseph, Renardy and Renardy (1983) are most important. The problem of stability is clearly identified as a problem of selection of stable arrangement of components in the latter paper. Interested readers should consult these papers and §10 of this paper for more references.

14. ENCAPSULATION INSTABILITIES. We say that a family of steady configurations of components in the flow of immiscible liquids undergoes an encapsulation instability when this family gives way to motions in which the high viscosity liquid is shielded from shearing by the low viscosity liquid. This type of instability can be observed as a migration of low viscosity liquid into regions of high shear. Sometimes the low viscosity liquid moves into the region of high shear as a sheet, and sometimes (for example, when the high viscosity liquid wets the moving boundary) the low viscosity liquid fingers into the high viscosity liquid, droplets are torn off the fingers and move into the region of high shear as an emulsion of droplets of low viscosity liquid in a high viscosity foam.

The series of experiments described below were motivated by observations of encapsulation instabilities in pipes. We wanted to know if the theoretical explanation of the observations which involved extremalizing dissipation could be defended and extended. We reasoned that minimum dissipation would put low viscosity liquid on a rotating rod, in the region where there is the greatest shear, as a kind of lubrication bonanza. Though our original idea was known by us to be oversimplified it was definitely useful as a guide to intuition and interpretation.

The experiments were carried out in two rectangular plexiglass boxes shown in Fig. 14.1. A box is filled with the heavier of two liquids up to the central diameter of the rod and the lighter liquid is floated on the top, as in Fig. 14.2. Since two liquids are always used the density differences are not vast. The rod is set into steady rotation. We want to know which liquid will coat the rod. The notion that the hydrodynamics will develop so as to minimize the torque, for a given speed, implies that low viscosity fluid will coat as in Fig. 14.3.

The experiments in which we achieved sheet coating of the type described by Fig. 14.3 are summarized in Table 14.1. Photographs of 5 entries from Table 14.1 are shown in Plates 14.1 - 14.5.

15. ROLLERS. There are very viscous oils which are extremely sticky in the sense that they strongly adhere to certain solid surfaces. Polymeric oils are sticky in this sense. High viscosity (μ = 950 poise) silicone oils and STP (μ = 110 poise) were sticky to aluminum and plexiglass in this sense. When the oils were floated on water the contact angle showed that aluminum favored water over STP (see Plate 15.1(a)) or silicone oil (water wets the rod). However, when the rod is turned on, it is the oil that coats, and in copious quantities. This is in apparent contradiction of the dissipation principle which, thinking superficially, would put water on the rod. So we have the impression that contact angles do not tell the whole story about stickiness.

Liquid 1	Liquid 2	μ_1 (Centipoise)	μ_1 (Centipoise)	ρ_1 (gm/cm^3)	ρ_2 (gm/cm^3)	Rod Wetted By:
silicone I	glycerine	19	1761	0.96	1.25	oil
silicone I	water	19	1	0.96	1.0	oil
light machine oil	glycerine	6.36	1761	0.831	1.25	oil
light machine oil	corn syrup	6.36	14	0.831	1.20	oil
silicone I	corn syrup	19	12	.96	1.20	oil
veg. oil	glycerine	60	1761	0.92	1.25	oil
castor oil	glycerine	700	1761	0.96	1.25	oil
water	dibutyl phthalate	~1	18-19	1	1.045	dibutyl phthalate
2% polyacrylamide	dibutyl phthalate	~700	18-19	1.02	1.045	dibutyl phthalate (sheet coating at high speed)
castor oil	1% polyacrylamide	700	90	0.96	1.01	oil (sheet coating at high speed)
silicone III	water	95000	1	0.95	1	water

Table 14.1

Experiments in which the low viscosity fluid coats in sheets.

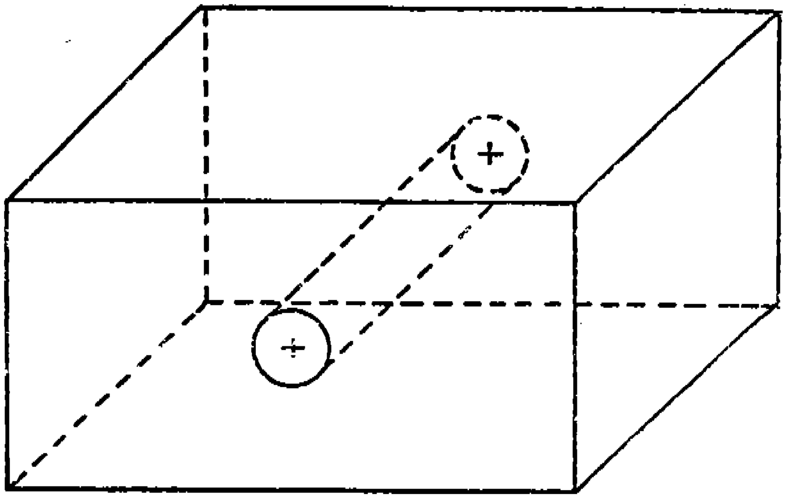


Fig. 14.1: Sketch of plexiglass boxes used in the experiments. A rod is inserted through the long planar sides of the box and is attached to a variable speed motor. The [length, height, depth, rod diameter, rod composition] of the two boxes, I and II, respectively are:

[20.32 cm, 22.86 cm, 10.16 cm, 5.08 cm, plexiglass]

for I and

[20.32 cm, 11.43 cm, 7.62 cm, 2.54 cm, aluminum]

for II.

The rod rotates counter-clockwise in box I and clockwise in box II. The experiments in Plates 14.1 - 14.4, 15.3, 16.1, 16.2 and 17.1 are carried out in box I and the experiments in Plates 14.5, 14.6, 15.1, 15.2, 16.3 and 16.4 are carried out in box II.

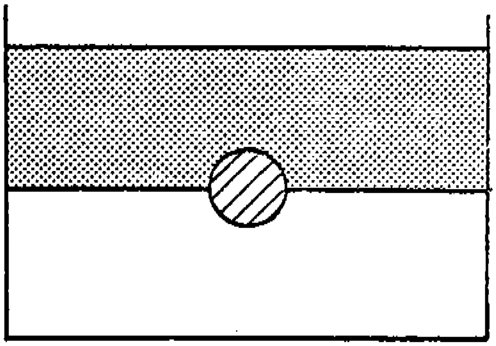


Fig. 14.2: The plexiglass box is loaded with two liquids with the undisturbed interface at the level of the horizontal diameter of the rod.

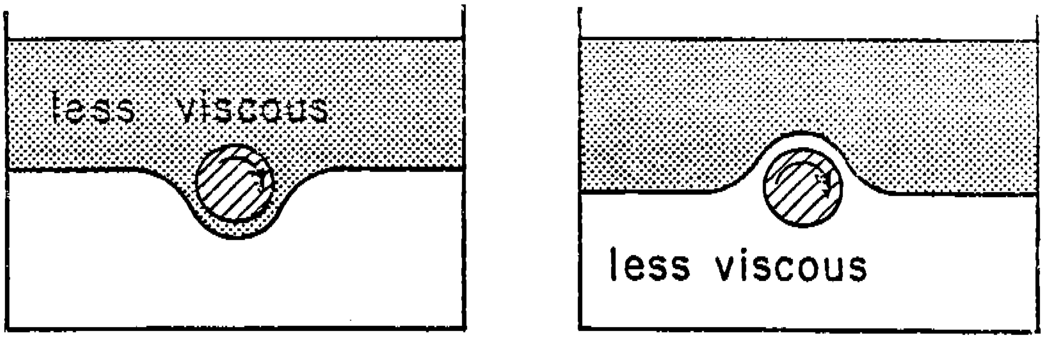


Fig. 14.3: The rod rotates and the less viscous liquid coats the rod. This kind of coating is called sheet coating. We say that the low viscosity liquid encapsulates the high viscosity liquid because it shields the more viscous liquid from intense shearing.

For the very sticky and viscous oils the configuration which appears to minimize dissipation is sometimes achieved by a kind of miracle which we call rollers. In Plate 15.1(a) we show the static configuration in which STP is floated on water. The STP wets the plexiglass box and water wets the aluminum rod. The rod was put into steady rotation (clockwise in Plates 15.1); it transported all of the STP on the left side of the box (i.e. the upward-motion side of the rod) to the roller of STP, or to a stagnant region of STP on the right (i.e. the downward-motion side of the rod). A thin sheet of water was pulled between the STP roller and the stagnant STP (see Plate 15.1(b) and Fig. 15.1). Astonishingly the roller of STP also separated from the wall of the box even though it is well known to us that STP is very sticky to plexiglass. This hydrodynamically generated separation of STP from the plexiglass side wall is shown in Plate 15.1(c).

The STP roller rotates as a rigid wheel, lubricated by water from all sides. The stagnant STP on the right barely moves. A sketch of Plate 15.1 is shown in Fig. 15.1 so that there is no ambiguity about what is being shown. This hydrodynamic configuration evidently reduces the total dissipation to a very small value associated mainly with shearing water in a lubrication layer.

Data for experiments leading to rollers is tabulated in Table 15.1.

The rollers are not hard to obtain. They seem to arise out of an encapsulation instability in which the water spontaneously migrates into the regions in which it undergoes high shear, shielding the STP from intense shearing. We have not yet studied the rollers under systematic variations of the parameters. However, in another set of experiments we tried to remove some of the STP from the stagnant region. In fact we removed all but about a 6mm layer of this STP without visibly affecting the stability of the roller. But at a critical value of the depth of "dead" STP, the roller bifurcated into another roller with triangular symmetry. This bizarre triangular "figure of equilibrium" was unstable but not violently so. In an attempt to save the day we added some STP to the stagnant region and recovered stability. However, the newly stable roller was lop-sided and ugly so we put a screwdriver in the box and molded it, as does a potter at his wheel, into the automobile tire shape of large radius which is exhibited in Plates 15.2(a) and 15.2(b).

The STP rollers are robustly stable. They withstand large perturbations and can rotate for weeks without apparent change.

The development of rollers might be thought to be associated to a degree with normal stresses characteristic of shear flows of non-Newtonian fluids. We discount explanations of our observations based on normal stresses because the shear rates in the rollers appear to be small and because we can obtain rollers in Newtonian liquids.

The photographs exhibited in Plates 15.3 show rollers of silicone oil (950 poise) in water. The viscosity ratio is 95,000. The roller of silicone oil is almost perfectly round, and it has detached from the walls under hydrodynamic action. The angular velocity of the silicone roller is constant; the roller rotates as a rigid wheel. The roller is robustly stable, it rotates for weeks without change of form. The dynamics of the roller appears to be governed by the inviscid equations of motion. There is an unknown constant in the pressure which determines the radius of the roller, through the balance of normal stress with surface tension. This radius

Liquid 1	Liquid 2	μ_1 Centipoise	μ_2 Centi-poise	ρ_1 g/cm^3	ρ_2 g/cm^3
STP	water	11000	1	0.89	1.0
Silicone II	water	120	1	0.95	1.0
Silicone III	water	95000	1	0.95	1.0*

*roller is formed only as a transition and the roller is lop-sided

Table 15.1

Experiments in which the high viscosity liquid formed rollers lubricated by water.

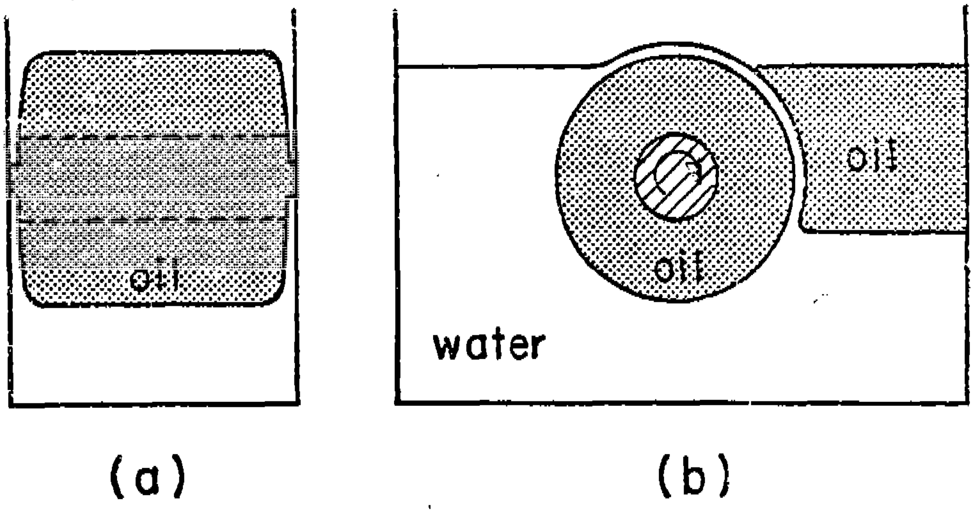


Fig. 15.1: (a) Sketch of plate 15.1(c); (b) sketch of plate 15.1(b).

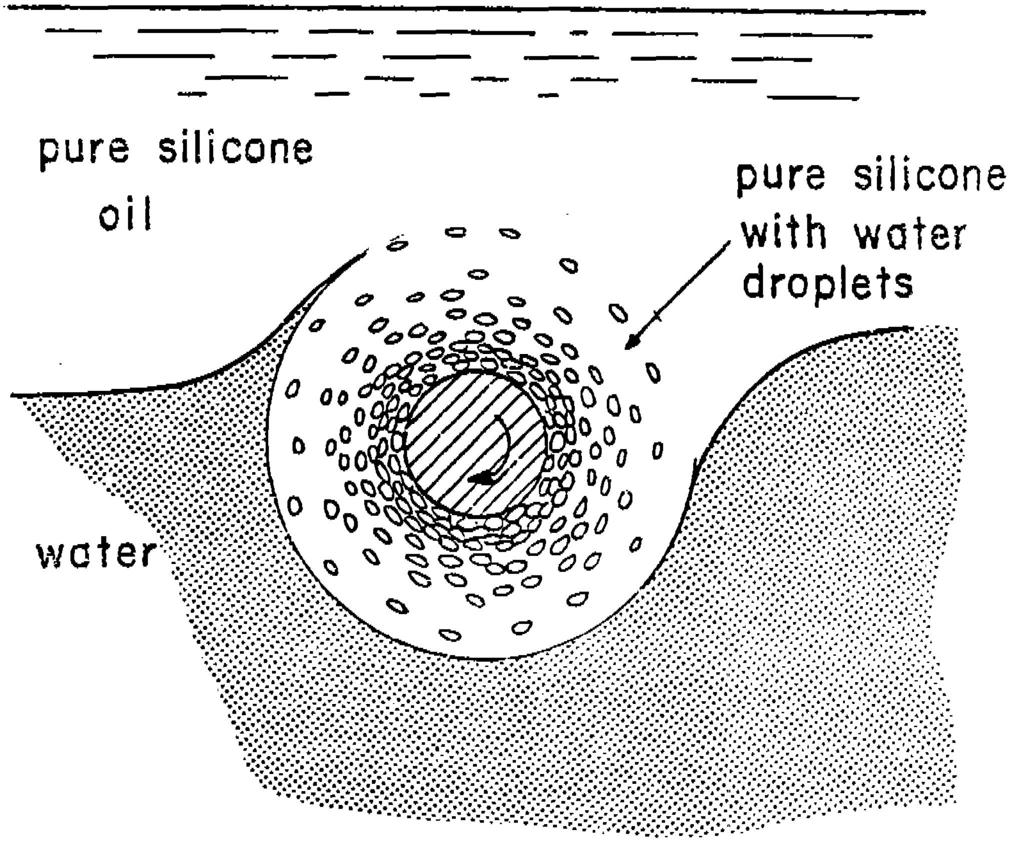


Fig. 15.2: Fingering of water droplets into high viscosity (95000cp) silicone oil.

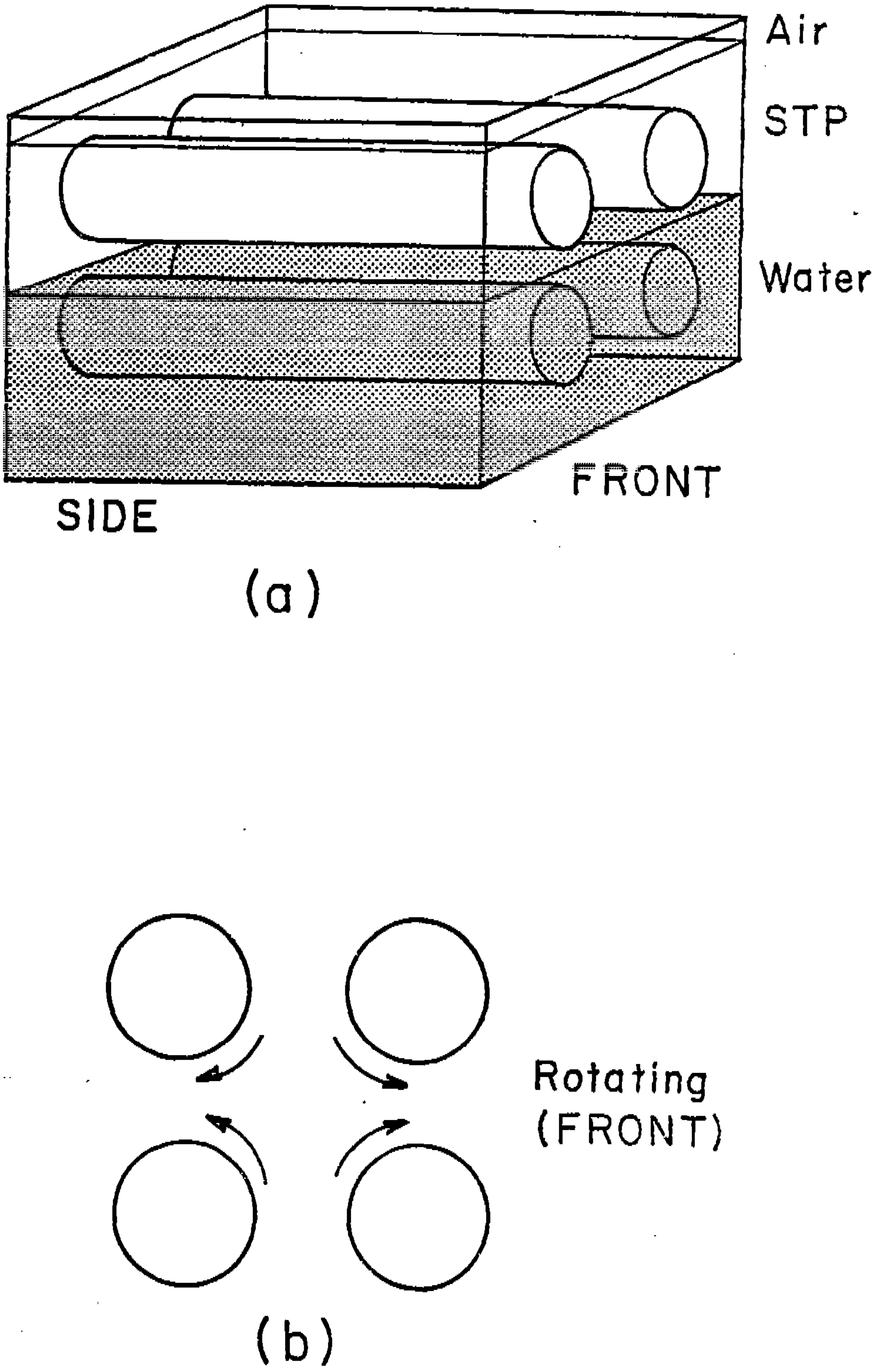


Fig. 15.3: Schematic of the experiment with STP and water in the four-roller apparatus.

is not uniquely determined. In fact the roller in Plates 15.3 has captured all of the silicone oil which was originally floated on the water. Presumably if we had floated more or less oil we would achieve a roller with larger or smaller radius. The silicone oil is attracted energetically to the plexiglass rod; it is favored over water.

We next describe what happened when we tried to obtain rollers with the same silicone oil (μ = 95,000 cp) in the box with an aluminum rod which was wetted by water rather than oil. At first, over a period of about six hours, we developed an imperfect roller configuration but thereafter we lost the lubricating water film to a cusp-like water film which is characteristic of fingering instabilities and the formation of emulsions (see Fig. 15.2). The configuration of components then assumed form as a dilute water laden emulsion of water droplets in silicone oil. The water droplets were continuously generated from the cusp shown in Fig. 15.2. The droplets were very effective in reducing the viscosity of the silicone oil. More and more of the water droplets drifted to the rod. After a few days this collection of drops reached a percolation threshold with rings, like wedding bands, of water around the rod. After three days all the rings had collected into a sheet of water coating the rod and encapsulating the silicone oil. The silicone oil in this configuration appears not to move, though of course there must be some small motion of the silicone oil due to shearing by water. This final configuration is shown in Plate 14.4.

The hydrodynamical history of the silicone oil experiments exhibits encapsulation phenomena in the form of roller instabilities, fingering instabilities and the generation of emulsions and finally to an unambiguous sheet encapsulation of high viscosity silicone oil by a lubricating water layer.

The sheet encapsulation can actually be put into a more dramatic form. In this configuration we get a water layer on the rod, completely surrounded by silicone oil as in the idealized picture shown in Fig. 7.1. This was achieved by withdrawing some water with a syringe after the sheet coating with water, described in the previous paragraph, has completely stabilized. The fully encapsulated configuration, shown in Plate 14.5, is robustly stable.

Arrays of rollers of STP lubricated on all sides by water can be achieved using the four roller apparatus of G. I. Taylor sketched in Fig. 15.3. The apparatus is filled with water up to the plane of symmetry between the upper and lower pairs of rollers, and then STP is added to cover the upper rollers. Rollers of STP develop out of small sinusoidal disturbances of initially uniform (along rod generators) interfaces. The interpenetrating rollers which finally develop are steady, stable and lubricated everywhere by water. Photographs of these rollers are exhibited in Plates 15.4(a) and (b).

16. FINGERING INSTABILITIES, THE FORMATION OF AND LUBRICATION BY EMULSIONS.

In many cases the wetting properties of the rod or the experimental conditions do not allow the formation of lubricating sheets of low viscosity liquid. In these cases we get fingering of low viscosity liquid into high viscosity liquid. Drops of low viscosity liquid are torn off the finger tips leading to the formation of an emulsion of low viscosity drops in a high viscosity foam.

A type of capillary instability may be associated with drop formation from fingers. The emulsions have a very low viscosity and they shield the bulk of high viscosity liquid from shearing. In Table 16.1 we have summarized the experiments exhibiting fingering into emulsions.

The first group of emulsifying configurations are those in which a light liquid of moderately high viscosity, like vegetable oil, light machine oil or castor oil, is floated on top of water, or waterbased polymeric solutions, like polyacrylamide. In all these cases the oil wets the plexiglass rod, and after the whole rod is exposed to oil, the oil clings tenaciously to the rod in a narrow layer, even in a monolayer, in apparent but only superficial contradiction of the lubrication principle. When the rod rotates slowly there is a tendency for water to be drawn up onto the rod, but surface tension pulls the water back as shown in Plate 16.1(a). At a higher speed the water will begin to finger into the oil, depositing droplets as shown in Plate 16.1(b). The fingering instability is sketched in Fig. 16.1. The continuous formation of droplets leads eventually to emulsification of water droplets in oil foam. The emulsified liquid then coats the rod as shown in Plate 16.1(c). Instead of sheet coating we get coating by water-laden emulsions. These emulsions have very low viscosities; first they are water-laden, second, they tank tread like roller bearings and they seem to be nearly as effective as sheet coats in shielding the high viscosity liquids from shearing.

The photograph of castor oil above polyacrylamide shown in Plate 16.2 is a variant of fingering dynamics leading to drops and emulsions. In this, the polyacrylamide-water droplets are encapsulated at higher speed by a sheet of low viscosity (polyacrylamide-water) liquid.

A second group of emulsifying configurations are generated by experimental conditions which prevent the generation of sheet coating. We did some experiments in Box II of the type that led to the formation of the rollers shown in Plates 15.1(a), (b) and (c). The only difference was that the box was filled to the top and kept from moving there by a cover plate. We expected that the phase configuration of minimum dissipation would lead to the capture of low viscosity fluid on the rod, with most of it on the rod if the densities were nearly matched. The difference between this sequence of experiments and the ones in §15 is that the cover plate forces a kind of hydrodynamic lubrication at the top of the cylinder, promoting fingering and the formation of emulsions.

Realizations of the idea of the foregoing paragraph are shown in Plates 16.3(a) and (b). As always, we started with the static configuration of heavy fluid below as shown in Plate 15.1(a). In Plate 16.3(a) we see the phase configuration of STP (dark) and TLA 227 (light) after a few days of rotation. Both fluids are oil-based polymeric immiscible liquids. The density of TLA 227 is about 0.005 gm/cm^3 greater than the density of STP. It can be seen in Plate 16.3(a) that after a few days much of the STP had migrated to the rod, and the streamlines carrying in the STP from remote regions are evident. When the (dark) STP is drawn from remote regions to the rod it carries with it some (light) TLA 227 by shearing action. After a week the color differences were very faint and we were

Liquid 1	Liquid 2	μ_1 centi-poise	μ_2 centi-poise	ρ_1 gm/cm^3	ρ_2 gm/cm^3	Rod Wetted By:
veg. oil	water	60	1	0.92	1.0	oil
peanut oil	water	60	1	0.92	1.0	oil
castor oil	water	700	1	0.96	1.0	oil
light oil machine	water	6.36	1	0.831	1.0	oil
castor oil	1% poly-acrylamide	700	90	0.96	1.01	oil, emulsion at low speed
STP	TLA 227	11000	20000	0.89	0.895	STP
STP	1% poly-acrylamide	11000	90	0.89	1.01	polyacrylamide
STP	water	11000	1	0.89	1.0	water
2% poly-acrylamide	dibutyl phthalate	700	19	1.02	1.045	emulsion at low speed

Table 16.1

Experiments in which the low viscosity fluid coats by fingering in emulsions.

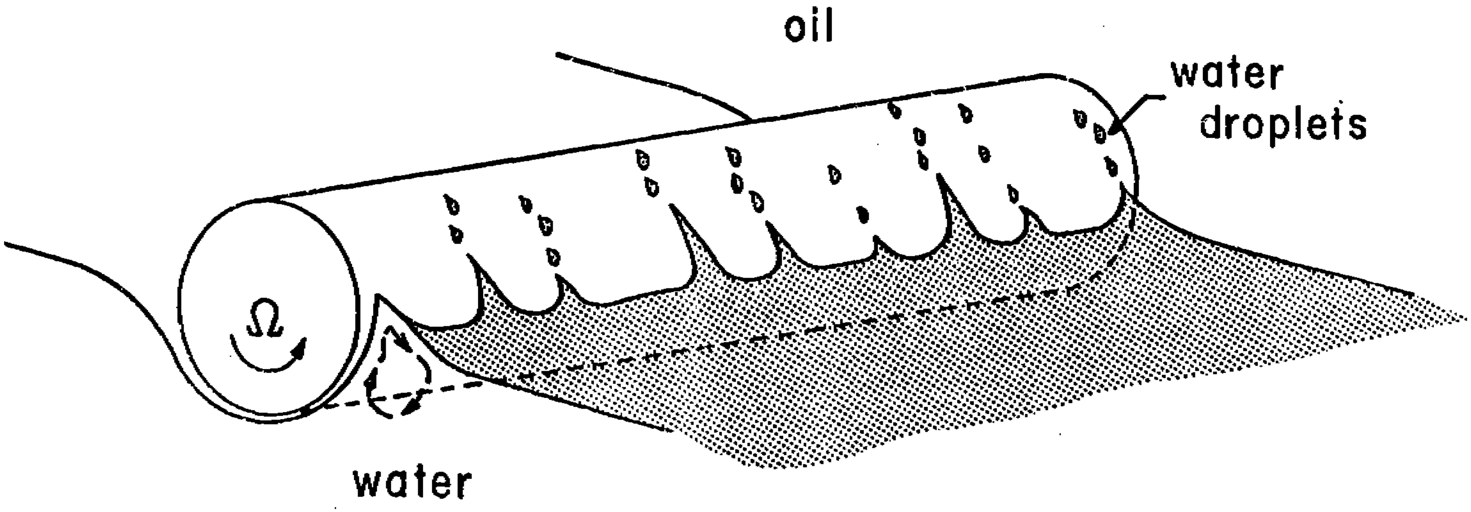


Fig. 16.1: Fingering instability leading to emulsions of water droplets in oil foam.

unsure if the bulk of the STP was really on the rod. But when we stopped the rod the STP precipitated out, as illustrated in Plate 16.3(b).

The shielding effects of emulsions is very clearly seen in the experiment with STP (dark) above polyacrylamide in water (clear), shown in Plates 16.4(a) and (b). The polyacrylamide is heavy and much less viscous (viscosity ratio of 120). The STP has a great affinity for the rod and when we turned on the rod it pulled along a big annulus of the high viscosity STP, going against our idea that the low viscosity fluid coats. But after a while the small bubbles of polyacrylamide which were being torn off were injected into the STP and the whole of the STP emulsified after about three days. The parts of the emulsion having a heavier concentration of low viscosity polyacrylamide drifted to the rod and all the shearing was confined to the low viscosity emulsion near the rod. Then the polyacrylamide in the emulsion near the rod deposited polyacrylamide in almost pure form onto the rod. This appears as a light ring of polyacrylamide on the rotating rod shown in Plate 16.4(b). The bulk of the STP is completely quiescent, being shielded from the polyacrylamide by the emulsion. We think that this configuration very nearly minimizes dissipation but we did not anticipate that the hydrodynamics would take on such bizarre forms. The rotational speed appeared to have increased a lot (at the same torque) in the week and one half of rod turning, suggesting a big drop in the dissipation of the preferred phase configuration.

17. CENTRIFUGAL AND TAYLOR TYPE INSTABILITIES IN THE FLOW OF IMMISCIBLE LIQUIDS. The instabilities we have in mind are the ones which are commonly associated with an adverse distribution of angular momentum. For single fluids such instabilities are well understood in certain circumscribed circumstances; Taylor cells are the best known and most important example. The point of novelty here is the presence of two liquids. Two types of phenomena which occur in our experiments are of interest. The encapsulation instability seems always to position a low viscosity film or emulsion between the rod and a stagnant body of high viscosity liquid. Such a configuration is very conducive to the development of an adverse distribution of angular momentum. The tendency for centrifugal forces to throw the heavy, less viscous liquid outward seems to be resisted by the other more viscous and still stable portion of the fluid. The result is apparently a cellular motion, which here can be observed as a deformation of the free surface. This is the first type of novel phenomenon which can be seen in experiments with two liquids. Such free surface cells are exhibited in Plate 17.1.

A second type of phenomenon develops in the flow of immiscible liquids in a Taylor apparatus set on its side. The inner diameter of the outer cylinder is 6.35 cm, and the length of the space between the inner and outer cylinders is 30.48 cm. The cylinders are made of plexiglass, the outer is stationary and the inner is free to rotate. Three inner cylinders (diameter = 5.72, 5.08, 3.81 cm) were used in the experiments. Every experiment was carried out with equal volumes of the two liquids. The apparatus was half-filled with heavy liquid on the bottom and light liquid on the top as in Fig. 17.1(a).

The fluid dynamics of the resulting flow is dominated by a form of Taylor instability which seems to be only weakly influenced by gravity. An idealized sketch of the cells which develop is shown in Fig. 17.1(b). High and low viscosity cells (e.g., oil and water) separate each other. Examples of the cells which actually do develop are exhibited in the photographs shown as Plates 17.2 - 17.6. Encapsulation instabilities enter into the dynamics of these cells in an important way which we specify below.

It is a good idea to compute stability limits for centrally located Couette flows of immiscible liquids in two layers with gravity neglected. This type of calculation would lead to critical conditions for the appearance of cells. For the present we imagine that there is a critical Taylor number, which is given by analysis for one fluid and is in the form

$$T_c = \frac{4\Omega^2 R_1^4}{\nu^2} \; \frac{\eta^2}{1-\eta^2} \; \frac{(1-\eta)^4}{\eta^4}$$

where R_1 is the radius of the inner cylinder, R_2 is the radius of the outer cylinder, $\eta = R_1/R_2$, ν is the kinematic viscosity and Ω is the angular velocity of the inner cylinder. For small $1 - \eta = \varepsilon$, $T_c = 3416$ and

$$3416 = \frac{2\Omega^2 R_1^4}{\nu^2} \varepsilon^3 .$$

In the experiments shown in the photographs exhibited in this paper the liquid with the smaller viscosity is water with $\nu = 0.01$ stokes. We may estimate $\varepsilon = 1 - \eta$ for the water alone as

$$\varepsilon = \tfrac{1}{2}\left(1 - \frac{R_1}{R_2}\right) .$$

Then, as a rough measure of critical conditions we have

$$\Omega_c = \frac{\sqrt{1708}}{100 R_1^2} \left[8\left(\frac{R_2}{R_2 - R_1}\right)^3 \right]^{\frac{1}{2}}$$

This is an estimate of the critical speed in water. The critical speed for oil is higher; roughly

$$20\Omega_c \text{ for silicone oil I } (\mu = 19\text{cp}),$$

$$12{,}000\Omega_c \text{ for STP } (\mu = 11000 \text{ cp}).$$

At low speed, we see at first a high torque, which very rapidly drops to a lower value. Correlating this with direct observation, we identify the first significant dynamical event as an encapsulation instability in which a sheet of water is pulled around the inner cylinder. After this event, gravity seems less important and the flows tend more to axisymmetry.

The next event is the formation of Taylor cells in the water layer. The dynamics associated with this event are not perfectly understood in detail but probably can be roughly described as follows. Imagine that a layer of water occupies the region next to the inner cylinder, with $[\![\rho]\!] = 0$ or, equivalently, $g = 0$. As Ω is increased past criticality the smooth flow of water gives up stability to Taylor vortices. The oil also moves very slowly in cells, driven not by instability but by shear stresses induced by cellular motion of the dynamically active water. These motions will naturally distort the oil/water interface, as in Fig. 17.2(a). The large amplitude limit of the flow in Fig. 17.2(a) is usually like the flow depicted in Fig. 17.2(c). In this flow the passive oil cells undergo extremely weak cellular motions driven by shears from active water cells. The oil cells are rollers and are lubricated at the sides and at the outer cylinder by water. The lubrication of oil rollers by water is enhanced by the fact that water is heavier than oil and will tend to replace the oil layer on the outside of the cylinder. If the oil is sticky to plexiglass a small layer of oil will continue to adhere to the outer cylinder at positions above the water layer where the shears are small. To understand this it is necessary to study Fig. 17.3. In the experiments it is very easy to see that the azimuthal component of velocity of oil cells near the outer cylinder is much greater than the azimuthal component of velocity in neighboring water cells. This striking observational fact is completely explained by the encapsulation of the oil rollers by water at the outer wall.

When the viscosity of the oil is not too great, it is possible to run our apparatus at speeds for which both oil and water are very unstable to Taylor vortices. In such situations the two components emulsify strongly, forming one emulsified liquid which exhibits Taylor cells in classical form (see Plate 17.3).

Acknowledgement

This work was supported by the U.S. Army Research Office and by the National Science Foundation.

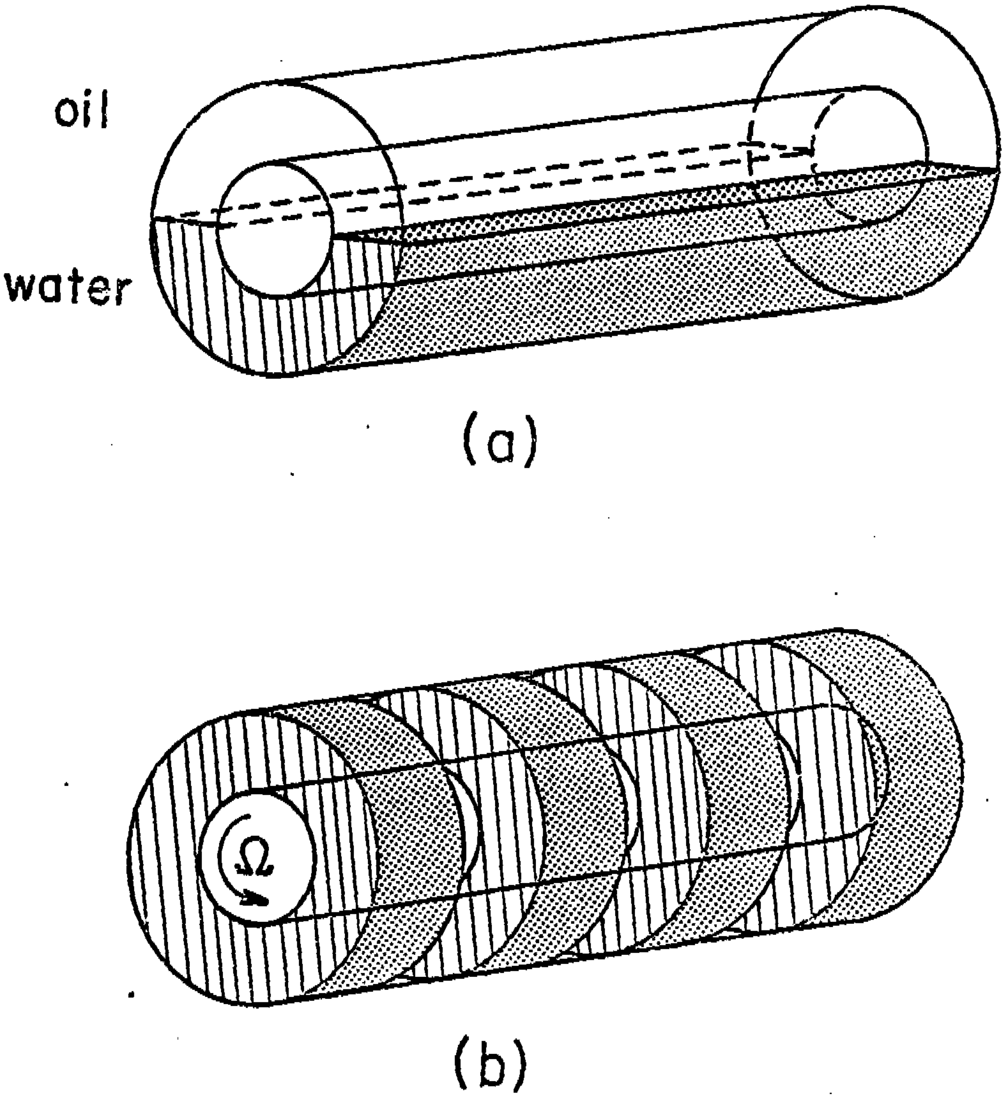


Fig. 17.1: Sketch of Taylor cells developing from the instability of water. The oil is dynamically passive while the motion in the water is driven by centrifugal instability.

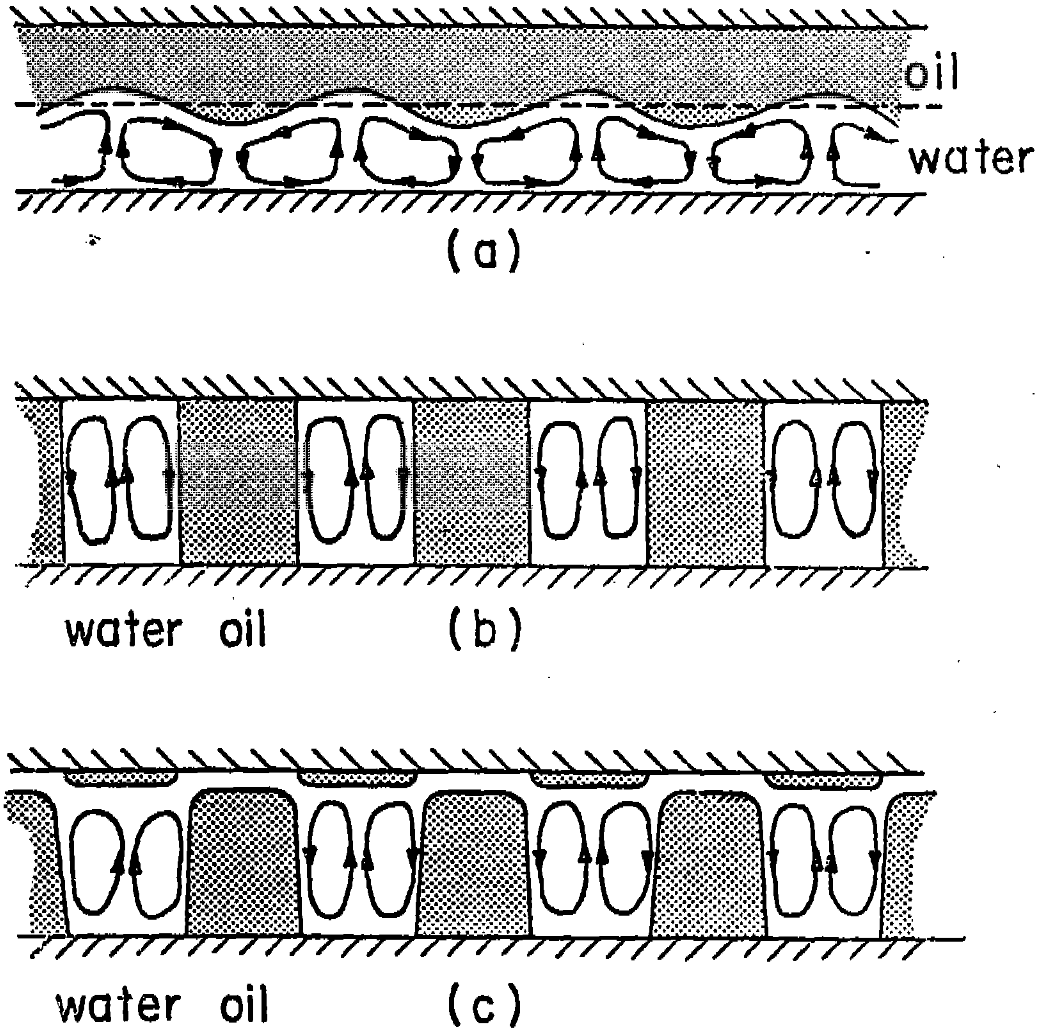


Fig. 17.2: Sketch of dynamically active water cells and dynamically passive oil cells (shaded) arising from the instability in bicomponent Couette flow between rotating cylinders. The situation in (a) arises near criticality as an instability of layered Couette flow. The situation in (b) could be regarded as the large amplitude limit of (a) when density differences are negligible. The secondary flow in the oil cells is extremely weak. Instead of (b) we usually see a configuration like (c) with passive oil cells rotating as rigid rollers attached to the inner cylinder and lubricated by water at the outer cylinder.

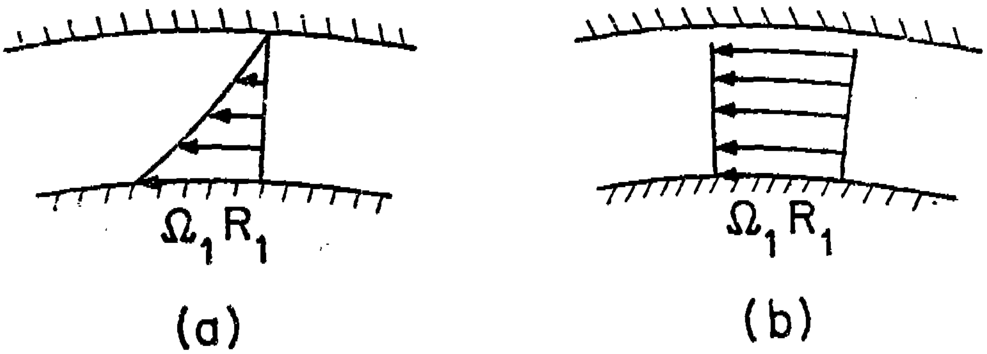


Fig. 17.3: The azimuthal velocity distribution in (a) water cells, and (b) oil rollers lubricated by water.

REFERENCES

Busse, F.H. 1982 Multiple Solutions for Convection in a two Component Fluid, Geophysical Research Letters 9(5), 519.

Charles, M.E. and Redberger, P.J. 1962 The Reduction of Pressure Gradients in Oil Pipelines by the Addition of Water: Numerical Analysis of Stratified Flow, Canadian J. Chem. Eng. 40, 70.

Everage, A.E. Jr. 1973 Theory of Stratified Bicomponent Flow of Polymer Melts. I. Equilibrium Newtonian Tube Flow, Trans. Soc. Rheol. 17:4, 629.

Hasson, D., Mann, U. and Nir, A. 1970 Annular Flow of Two Immiscible Liquids. I. Mechanisms, Canadian J. Chem. Eng. 48, 514.

Joseph, D.D. 1976 Stability of Fluid Motions II. Tracts in Natural Philosophy. Springer.

Joseph, D.D.,Renardy, M. and Renardy, Y. 1983 Instability of the Flow of Immiscible Liquids with Different Viscosities in a Pipe, to appear.

Lee, B.L. and White, J.L. 1974 An Experimental Study of Rheological Properties of Polymer Melts in Laminar Shear Flow and of Interface Deformation and Its Mechanisms in Two-Phase Stratified Flow, Trans. Soc. Rheol. 18:3, 467.

MacLean, D.L. 1973 A Theoretical Analysis of Bicomponent Flow and the Problem of Interface Shape, Trans. Soc. Rheol. 17:3, 385.

Minagawa, N. and White, J.L. 1975 Coextrusion of Unfilled and TiO_2-Filled Polyethylene: Influence of Viscosity and Die Cross-Section on Interface Shape, Poly. Eng. Sci. 15, 825.

Plateau, J.A.F. 1873 Statique Experimentale et Theorique Des Liquides. Gauthier-Villars, Paris.

Southern, J.H. and Ballman, R.L. 1973 Stratified Bicomponent Flow of Polymer Melts in a Tube, Appl. Poly. Sci. 20, 175.

Williams, M.C. 1975 Migration of Two Liquid Phases in Capillary Extrusion: An Energy Interpretation, AICHE J. 21(6), 1204.

Yih, C.S. 1967 Instability Due to Viscosity Stratification, J. Fluid Mech. 27, 337.

Yu, H.S. and Sparrow, E.M. 1967 Stratified Laminar Flow in Ducts of Arbitrary Shape, AICHE J. 13(1), 10.

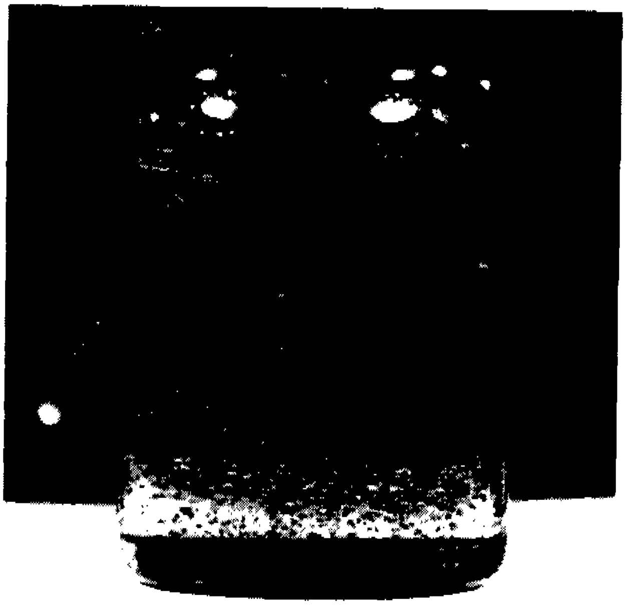


Plate 3.1: One of the static configurations of dibutyl phthalate bubbles (ρ = 1.04 gm/cc) in glycerine-water solution. The mixture of glycerine (ρ = 1.25 gm/cc) and water (ρ = 1.00 gm/cc) is adjusted to match the density of dibutyl phthalate.

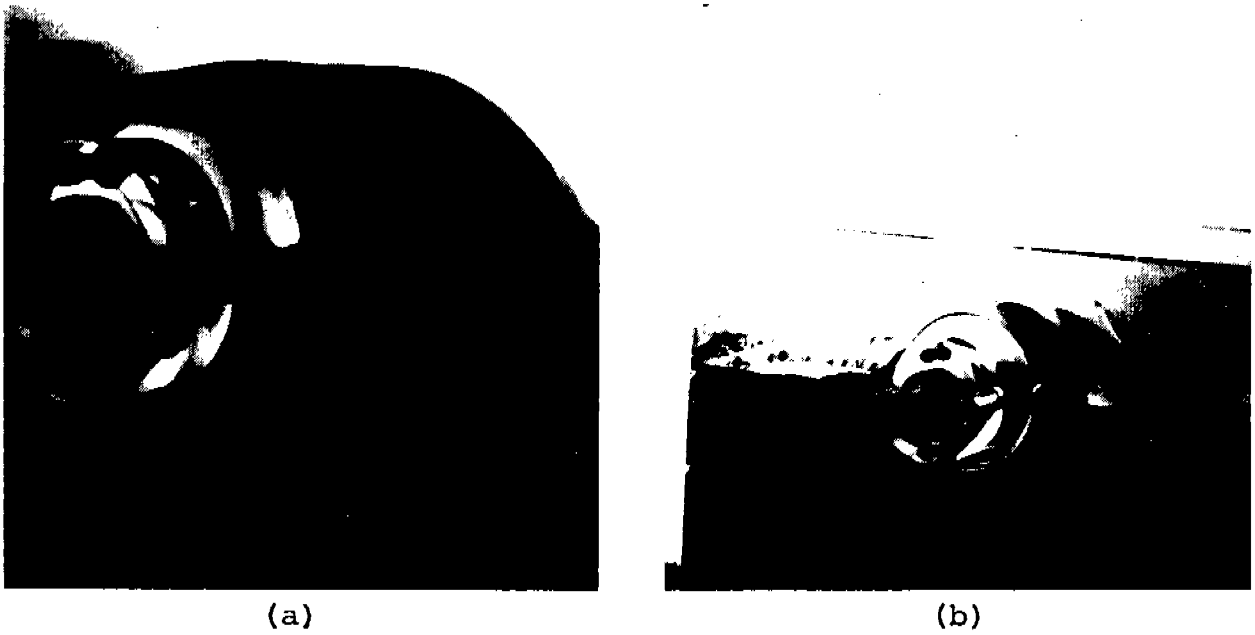


Plate 14.1: Encapsulation of silicone oil I (ρ = 0.96 gm/cc, μ = 19 cp) by (a) corn syrup/water solution (ρ = 1.2 gm/cc, μ ~ 12 cp, blue), and (b) water (blue). The speed of the rod is about 150 RPM.

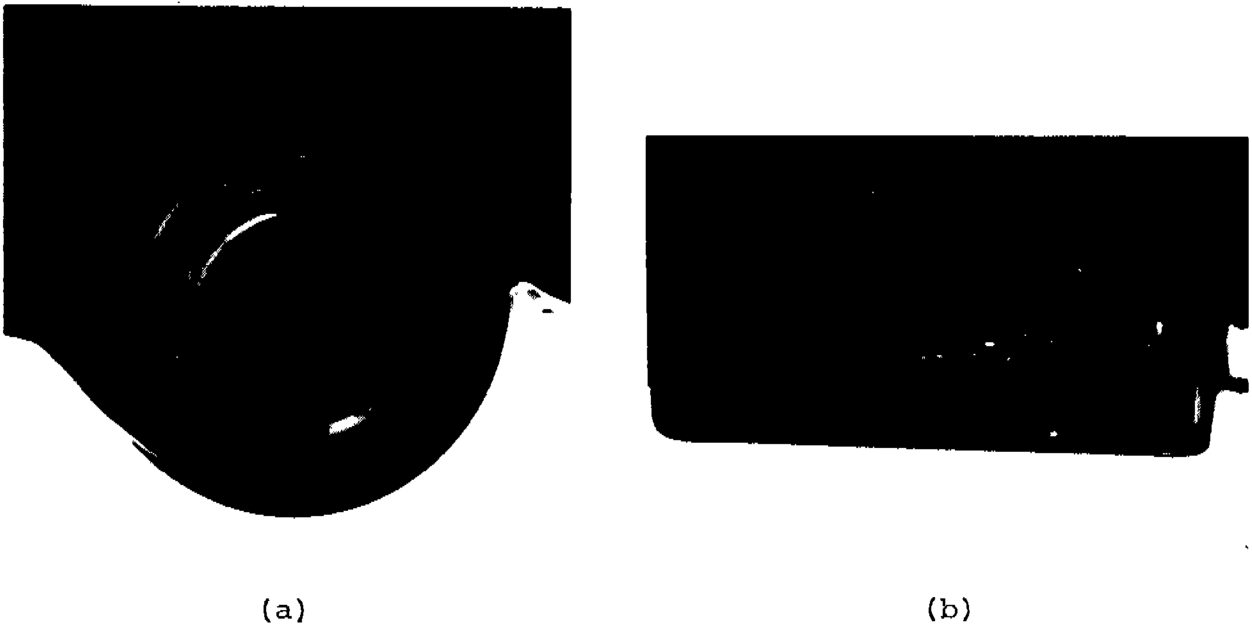
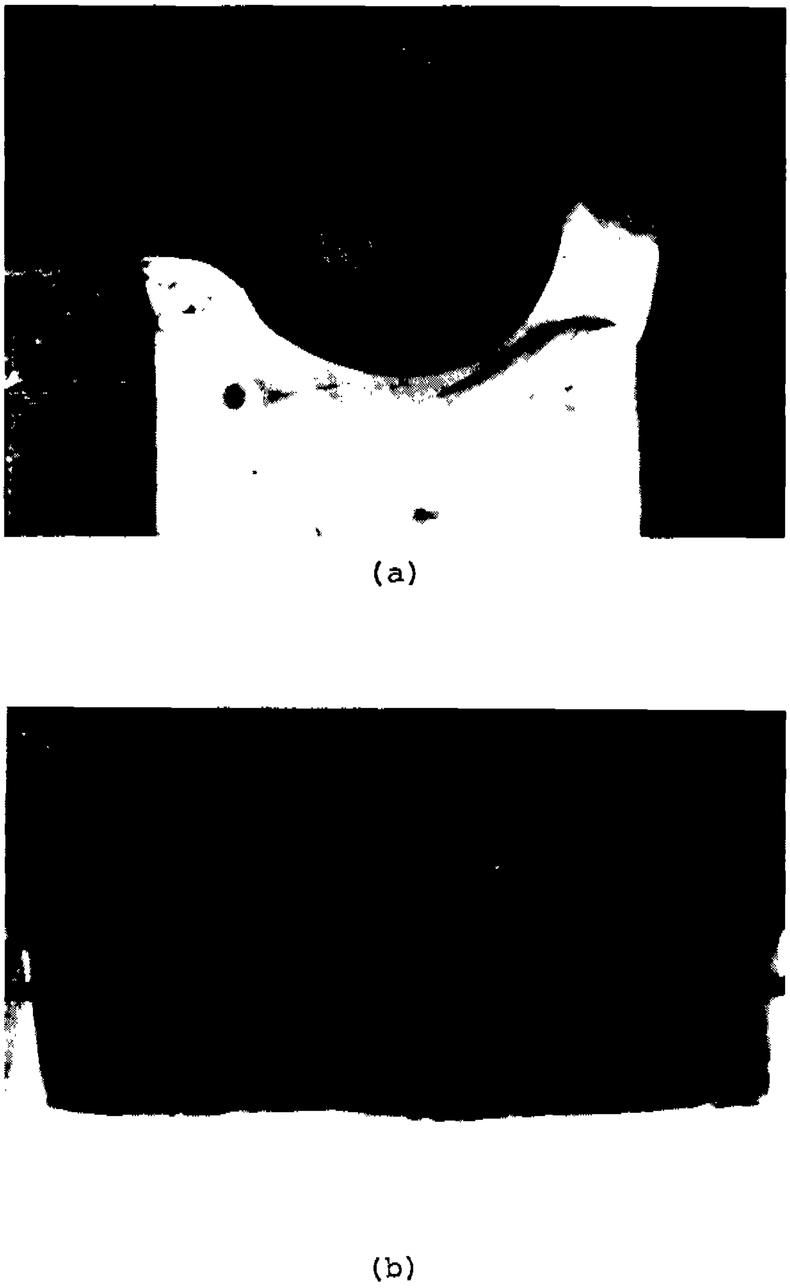


Plate 14.2: Encapsulation of glycerine (ρ = 1.25 gm/cc, μ = 1760 cp, clear) by silicone oil I (ρ = 0.96 g/cc, μ = 10 cp, red). The speed of the rod is 100 RPM: (a) front view, (b) side view.



(a)

(b)

Plate 14.3: Encapsulation of glycerine (ρ = 1.25 gm/cc, μ = 1760 cp, clear) by castor oil (ρ = 0.96 gm/cc, μ = 700 cp, red). The speed of the rod is about 50 RPM: (a) front view, (b) side view.

(a) (b)

Plate 14.4: Encapsulation of silicone oil III (ρ = 0.95 gm/cc, μ = 95,000 cp, red) by water (clear). The speed of the rod is about 65 RPM. The rod is lubricated entirely by water: (a) back view, looking at an angle from below, (b) side view.

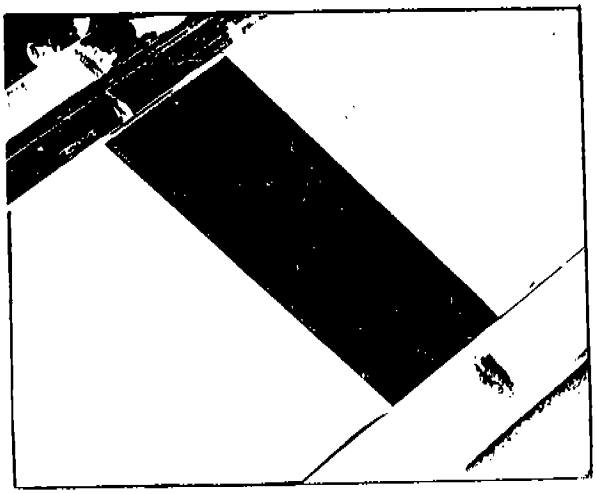


Plate 14.5: The same experiment as in Plate 14.4. After the configuration in Plate 14.4 is achieved, some water is withdrawn from the box so as to lower the level of silicone oil below the rod. The rod is surrounded by a very thin annular layer of water and the water is surrounded by oil. The rod is completely lubricated by the water and the oil is sheared only by water.

It is possible to obtain a water lubricated aluminum rod in which the thickness of the water layer is zero, confined to a monolayer on the rod. In such configurations the rod rotates at high speeds, but the silicone oil is dead still. This violates the no-slip condition; it slips, completely as in an inviscid fluid. So if we wish to say that fluid will stick to a solid we must specify the fluid, the solid, and say there are no monolayers, or make other quantifying statements.

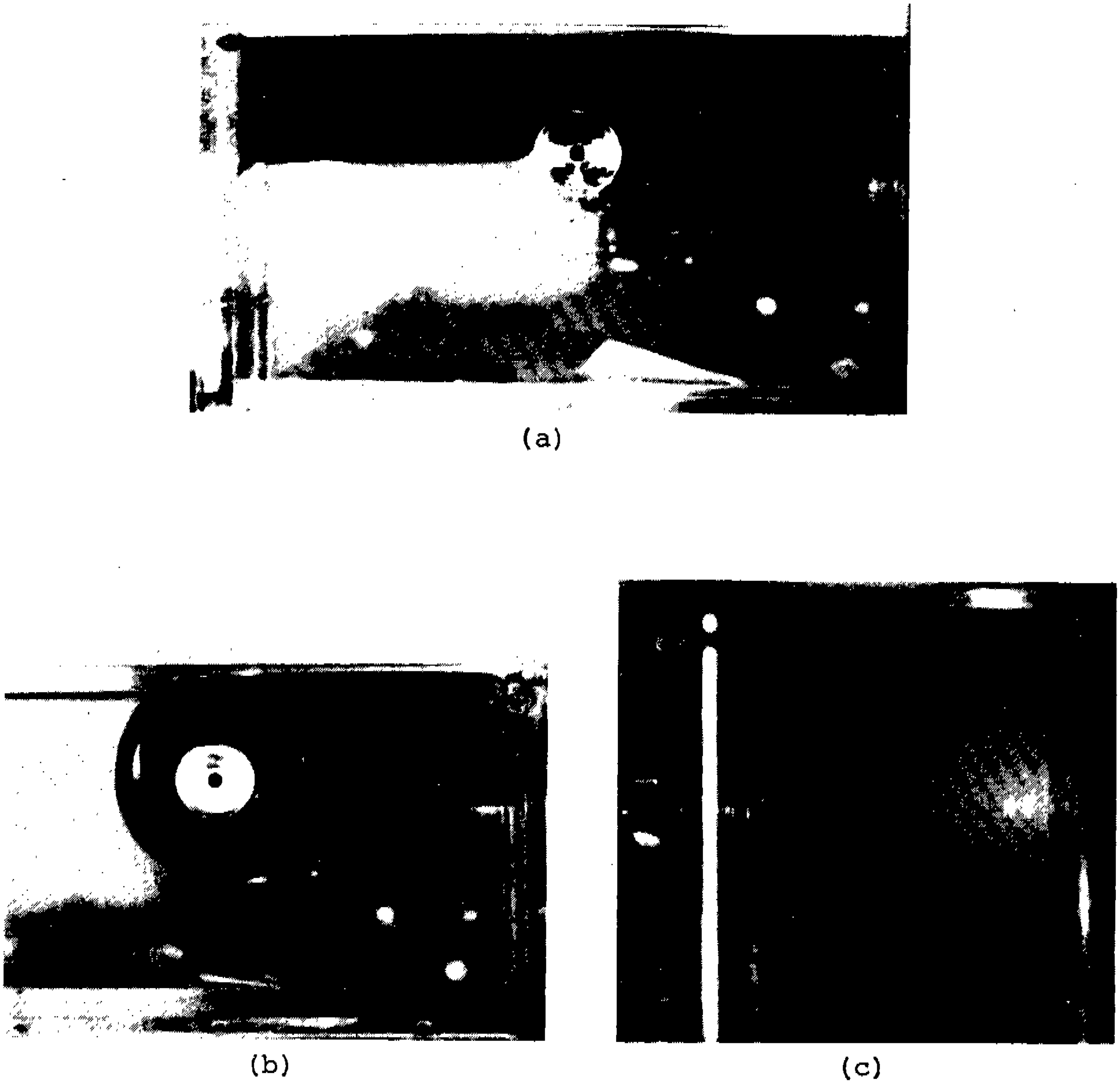
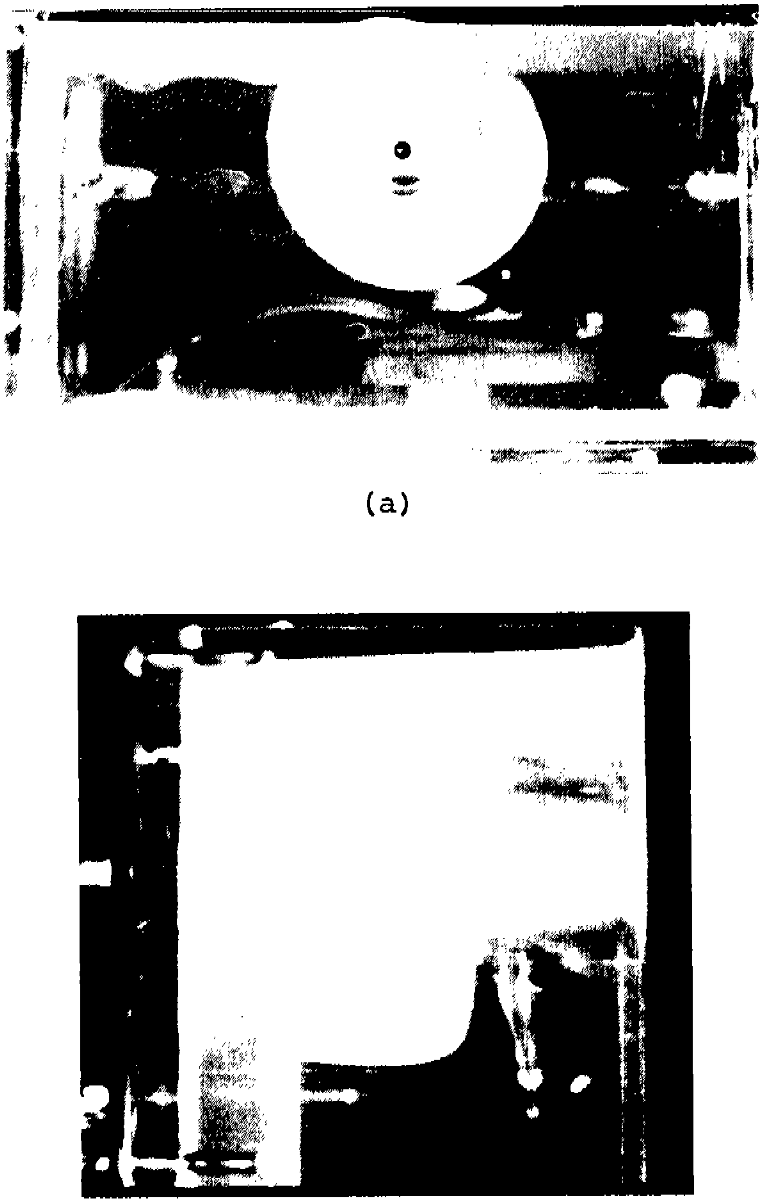


Plate 15.1: (a) The static configuration of STP (density $\rho = 0.89$ gm/cc, viscosity $\mu = 11000$ cp, brown) on water ($\mu = 1$ cp).

(b) Front view of the box. The rod rotates clockwise at about 40 RPM. The STP on the right is nearly stationary and is shielded from shearing by a thin layer of water.

(c) Side view of the box looking in from the left of Plate 15.1(b). The STP roller is also shielded from shearing against the plexiglass walls by a layer of water.



(a)

(b)

Plate 15.2: (a) The same experiment as in Plate 15.1(b) and (c) after some STP is removed from the right of the box, resulting in a roller in shape of an automobile tire. There is a smaller layer of STP on top of the water. This layer is separated from the roller by a layer of water maintained hydrodynamically.

(b) Side view of the roller, looking in from the left of Plate 15.2(a). The roller has detached from the side walls under hydrodynamic action.

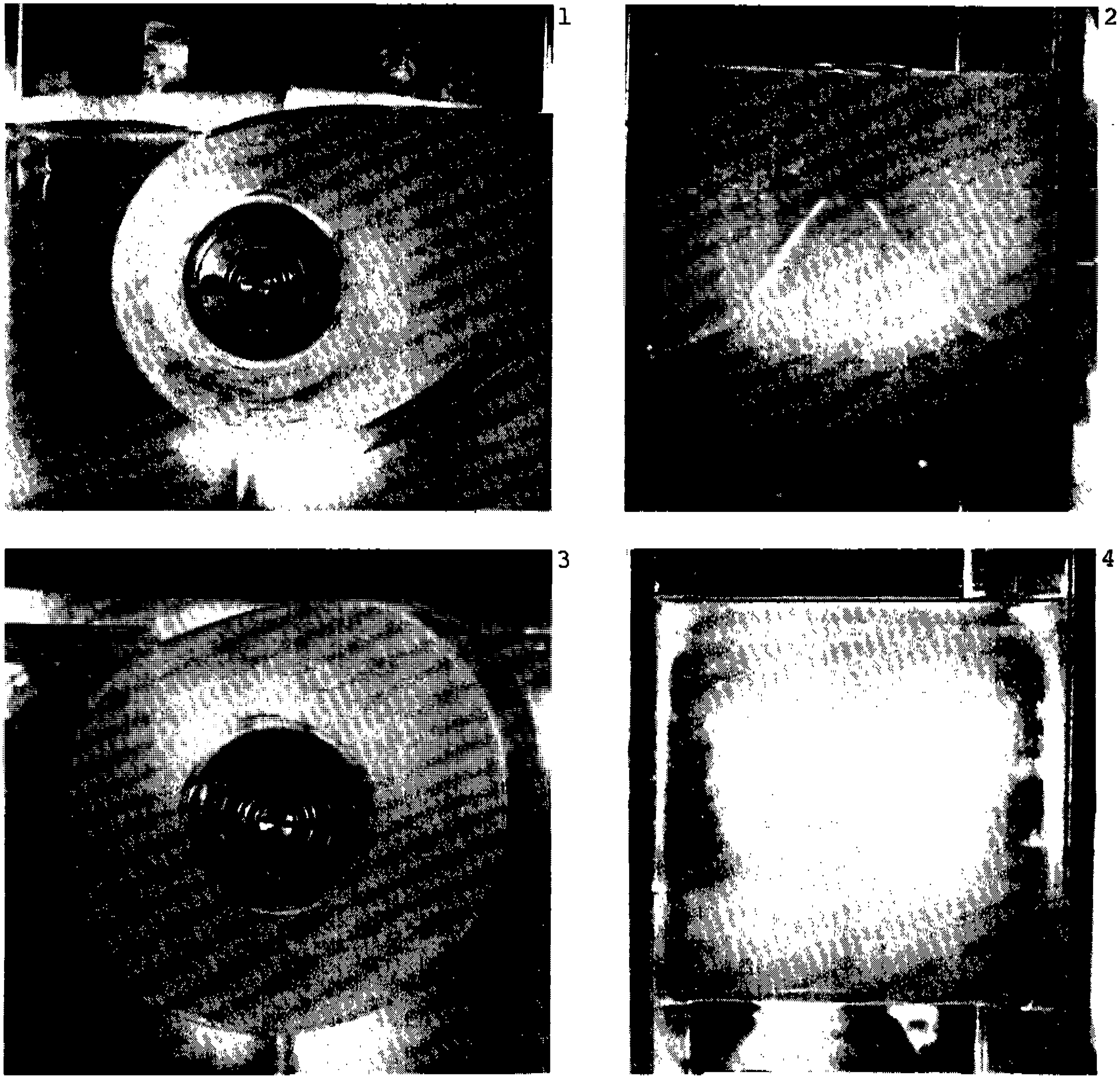
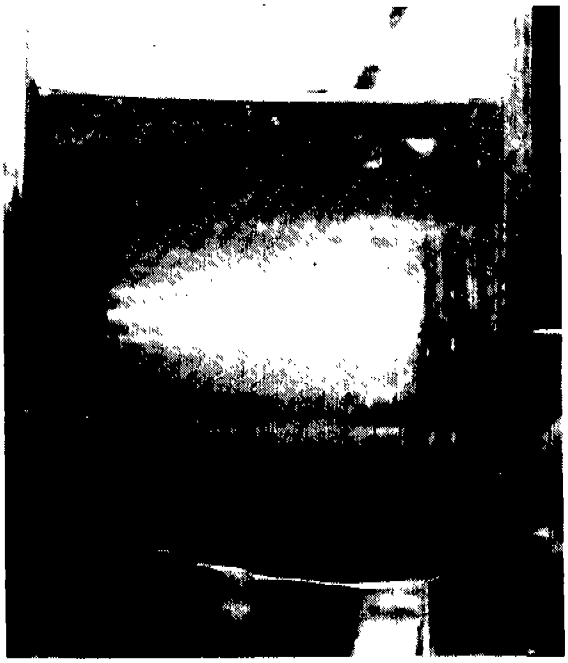
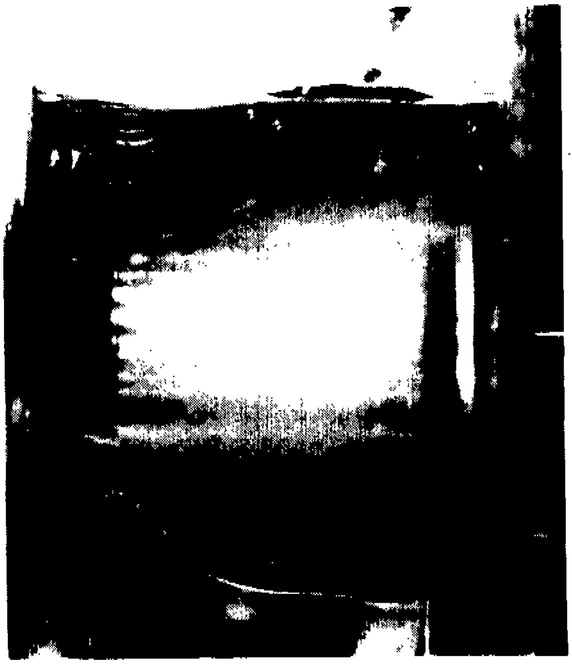


Plate 15.3: (a) The formation of a roller of silicone oil III ($\rho = 0.95$ gm/cc, $\mu = 95,000$ cp) in water. In the beginning the rod rotates counter-clockwise at 10 RPM. The torque is 1.4 lb.-in. Fingering instabilities lead to an emulsion of water droplets in silicone oil; (1) front view, (2) side view.

Three days later as more water droplets are formed, the effective viscosity of silicone oil decreases and the rod rotates at a higher speed (16 RPM). At this speed the roller is formed but does not rotate as a solid body since it is still attached to the side wall (3) front view. (4) side view.



1



2

Plate 15.3: (b) The speed of the rod is increased. The torque goes up to above 5 lb.-in. The two photographs (1) and (2), which are taken 5 seconds apart, show the dynamics through which the roller detached itself from the side wall. After detaching from the wall, the torque goes down to 0.6 lb.-in.

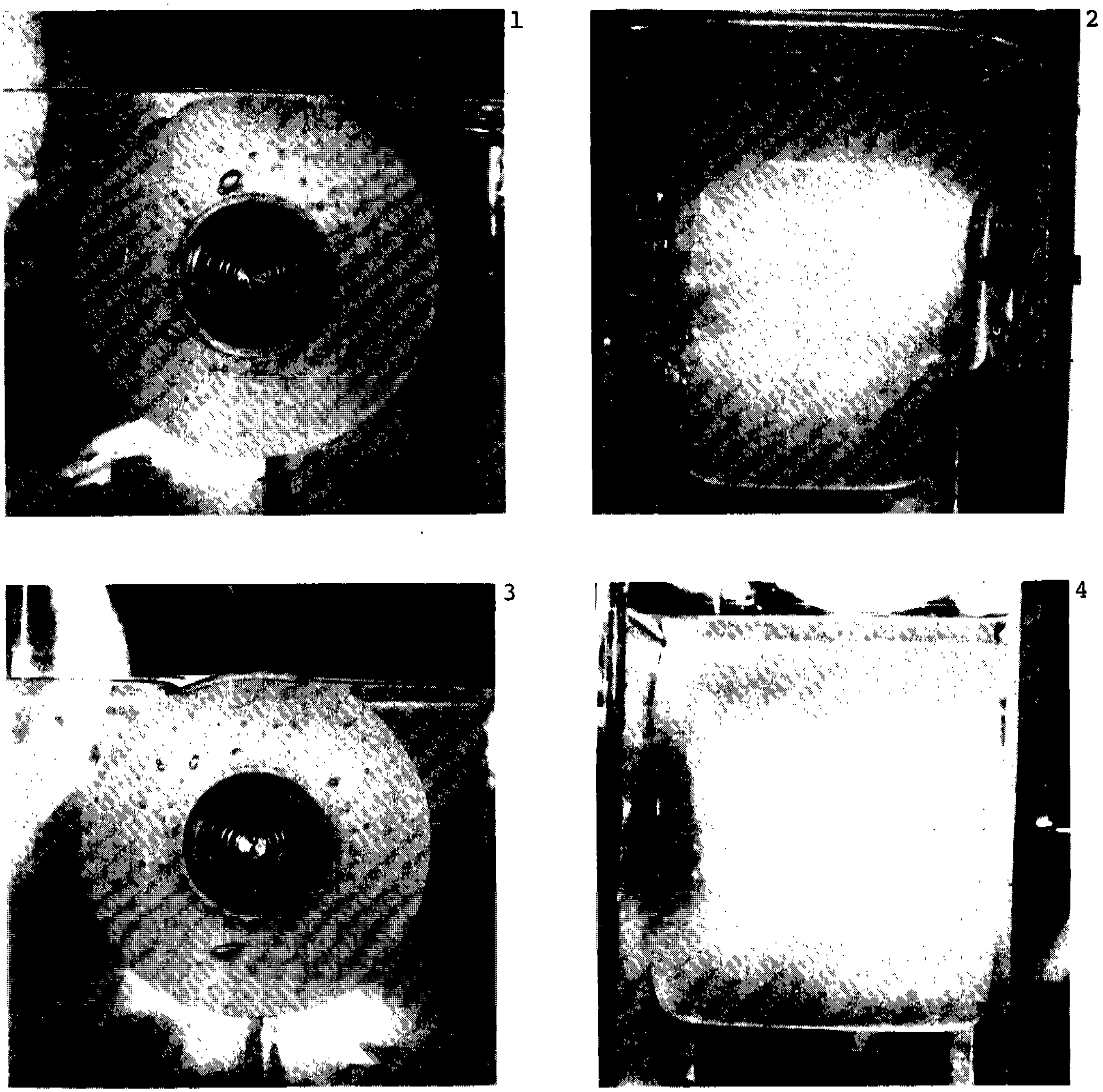
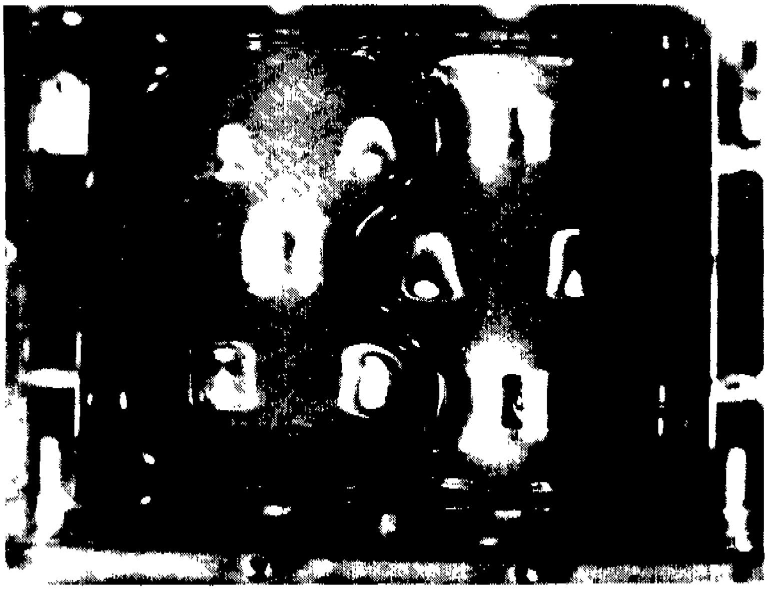
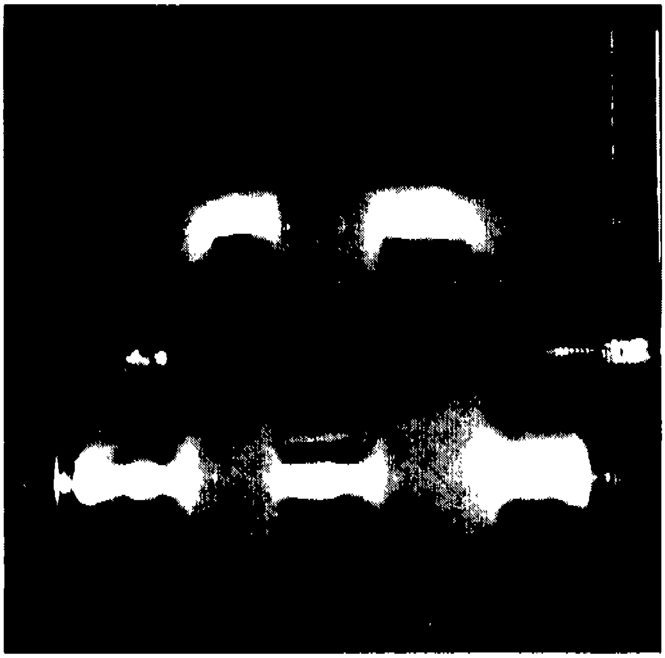


Plate 15.3: (c) After detaching from the wall the roller becomes unstable. To restore stability the speed of the rod is reduced to 12 RPM. The roller is stable but the shape is irregular; (1) front view, (2) side view. After a few hours the roller molded itself into the shape depicted in Fig. 15.1. The speed of the rod and the speed on the surface of the roller are almost the same showing that the roller is very nearly in a solid body rotation, with small shearing by water at the roller rim.



(a)



(b)

Plate 15.4: (a) Interpenetrating rollers of STP and water as seen from the top of the box sketched in Fig. 15.3. There is water on every side of the STP rollers. The rollers develop as an instability of two rollers (one on each rod) which are initially uniform along the rod. The water surface between them develops a small wave. A grown-up version of small waviness can be seen in the above photograph.

(b) Interpenetrating rollers of STP and water as seen from the side of the box sketched in Fig. 15.3. The clear parts are of water.

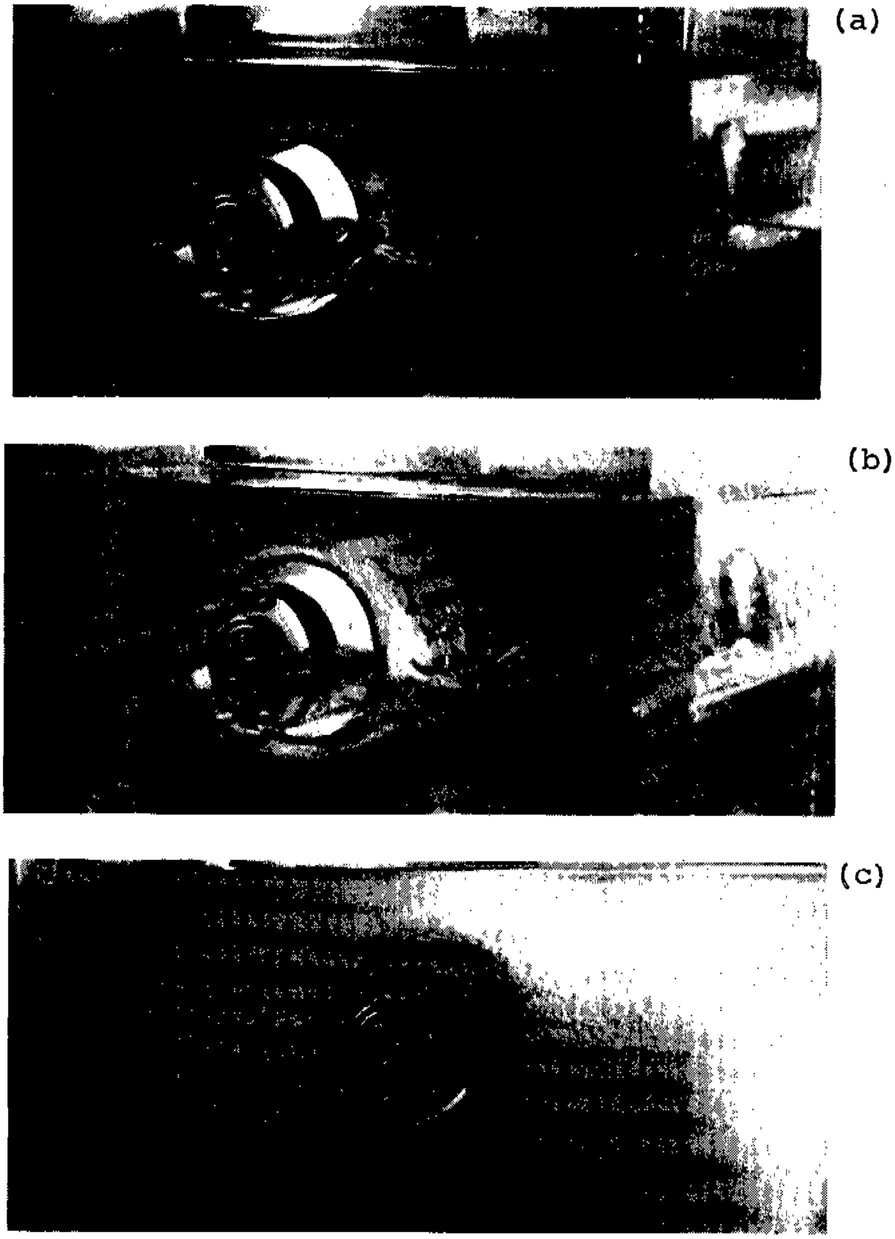


Plate 16.1: (a) Light machine oil (ρ = 0,831 gm/cc, μ = 6.36 cp, yellow) on water. The rod rotates counterclockwise at 115 RPM. The contact angle in this experiment seems to be fixed with the contact line slipping on the rod in such a way as to stay fixed in space. The configuration of the contact line and the tenacity of the contact angle even under pressure from the intense water circulation under the free surface are noteworthy.

(b) The rod rotates at 195 RPM. This plate shows the fingering instability in mature form and how water droplets are torn off the fingers.

(c) Emulsion of water in light machine oil at 300 RPM. The water-laden emulsion shields the main body of oil (on the top left and right) from shearing.

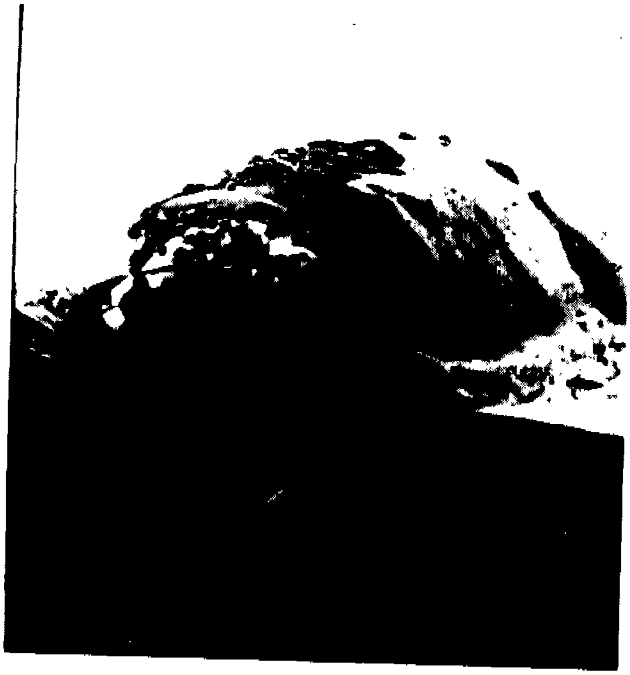
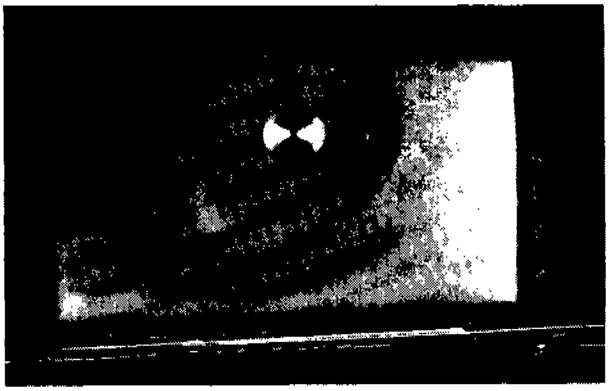
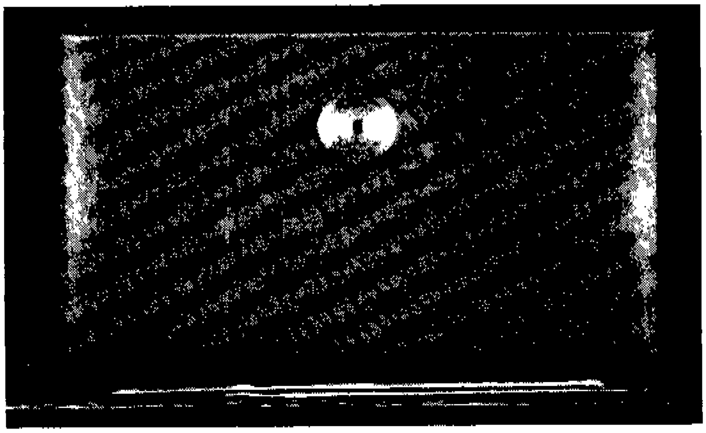


Plate 16.2: A combination of emulsion and sheet coating of 1% polyacrylamide/water (μ = 90 cp, ρ = 1.01 gm/cc, blue) in castor oil (μ = 700 cp, ρ = 0.96 gm/cc, yellow). The rod rotates clockwise at 1 RPM. The rod is covered by polyacrylamide/water droplets and these droplets in turn are encapsulated by a sheet of polyacrylamide/water shielding them from shearing against the high-viscosity castor oil.



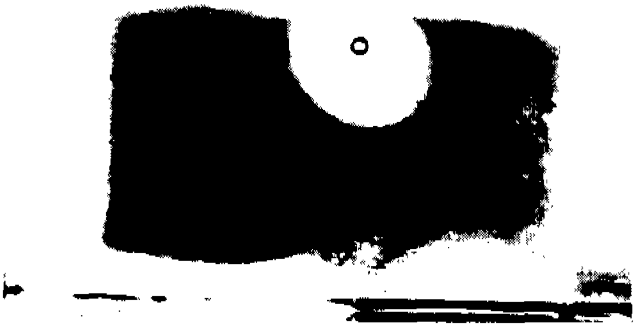
(a)



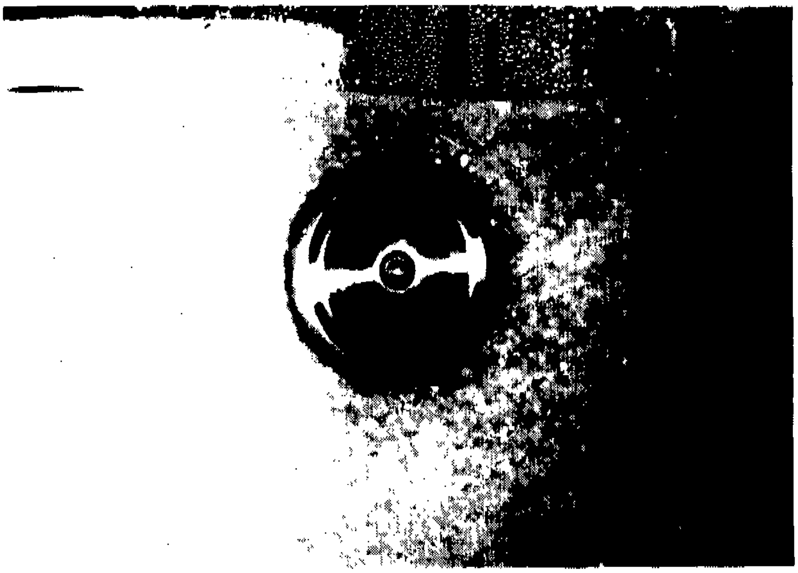
(b)

Plate 16.3: (a) Emulsion of STP (μ = 11000 cp, ρ = 0.890 gm/cc, dark brown) in TLA 227 (μ = 20000 cp, ρ = 0.895 gm/cc, light brown). The rod rotates clockwise at about 16 RPM. Streamlines show STP being drawn into the rod.

(b) Emulsion of STP in TLA 227. The motor is stopped. A ring of STP (dark brown) precipitates out of TLA (light brown) around the rod.



(a)



(b)

Plate 16.4: (a) Emulsion of 1% polyacrylamide/water (μ = 90 cp, ρ = 1.01 gm/cc) in STP (μ = 11000 cp, ρ = 0.890 g/cc, dark). The polyacrylamide/water solution is on the bottom and on the rod, which rotates clockwise at about 45 RPM.

(b) Close-up view showing the emulsion and the ring of polyacrylamide/water solution around the rod.

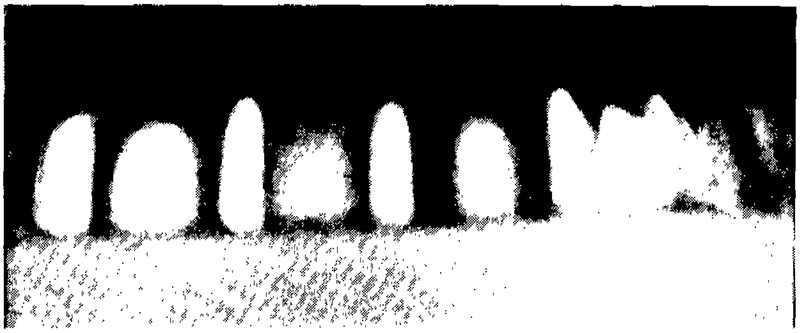


Plate 17.1: Centrifugal instability of multigrade motor oil (10W40) in water. The layer of oil on the rod is separated from the main body of oil by a sheet of water. At the speed of about 300 RPM the oil-water interface becomes wavy.

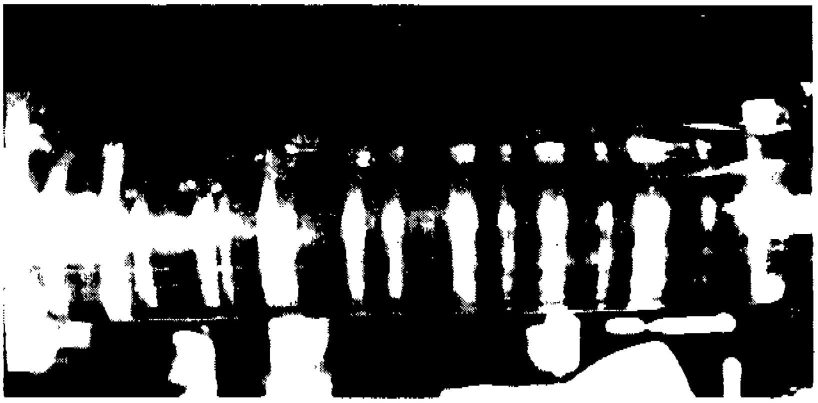


Plate 17.2: Rollers of silicone oil I ($\rho = 0.95$ gm/cc, $\mu = 19$ cp, red) separating Taylor cells of water (clear) $(\Omega, R_2 - R_1) = (130$ RPM, 0.635 cm$)$. The azimuthal velocity of silicone oil is much larger than water, presumably because of encapsulation by a thin film of water (see Fig. 17.3).

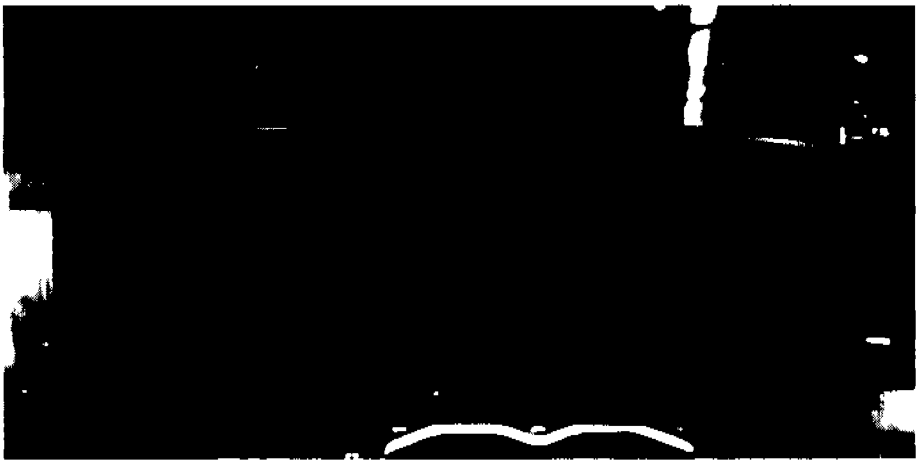


Plate 17.3: A similar experiment to Plate 17.2 at a much higher angular velocity, $(\Omega, R_2 - R_1) = (1810$ RPM, 0.318 cm$)$. At this velocity the oil and water are both unstable and the oil has completely emulsified, forming a single liquid in which we see classical Taylor cells.

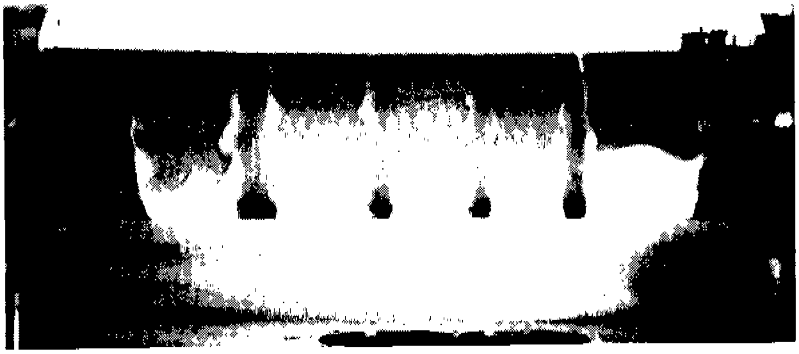


Plate 17.4: Rollers of STP (orange) separating Taylor cells of water (clear); $(\Omega, R_2 - R_1)$ = (86 RPM, 1.27 cm). The viscosity of the STP is 11,000 that of water, so that STP is always stable against Taylor instability. The azimuthal velocity of the rollers is much greater than that of the wave of STP which sticks to the inner wall of the outside cylinder above the water (see Fig. 17.2(c) and 17.3). The manner in which the STP is fractured to avoid being sheared is noteworhty.

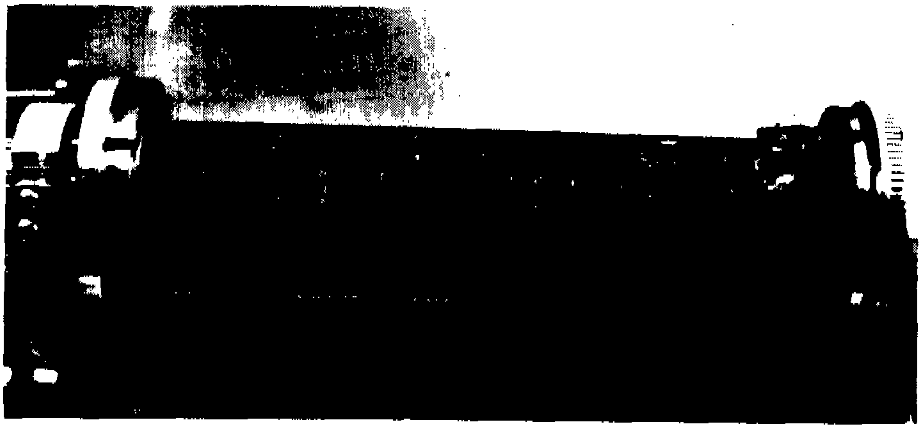


Plate 17.5: Monograde motor oil (SAE40, brown) and water (clear), (Ω, $R_2 - R_1$) = (120 RPM, 0.318 cm). At this speed only water is unstable, and the dynamics is similar to (c) of Fig. 17.2. We see rollers of oil separating Taylor cells of water. The azimuthal velocity of water is much smaller that that of oil, apparently caused by the layers of oil sticking to the inner wall of the outer cylinder.

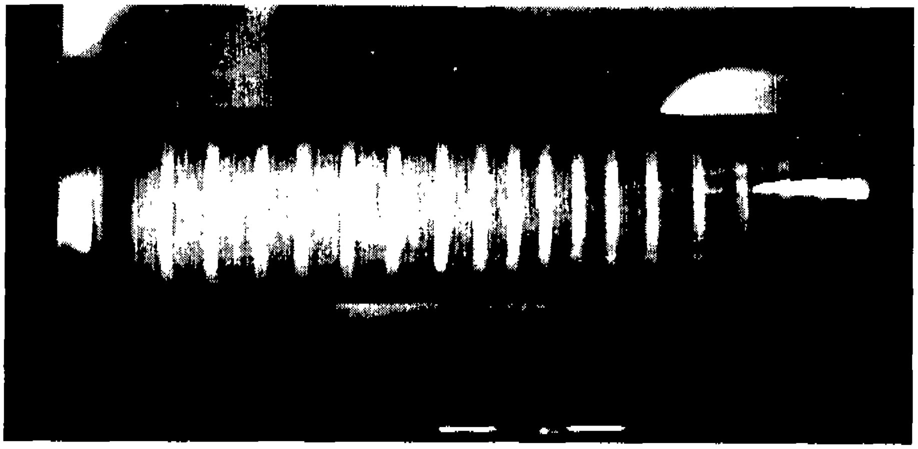


Plate 17.6: Monograde motor oil (SAE40) and water (Ω, $R_2 - R_1$) = (440 RPM, 0.318 cm). The light cells are emulsions of oil with small droplets of water. The light cells are of water with many large drops of oil. This is like a two component flow of two different emulsions, and the dynamics which are realized seem to fall under (b) of Fig. 17.2.

FLOW PAST A FLEXIBLE MEMBRANE

Jean-Marc Vanden-Broeck
Department of Mathematics and
Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53706

ABSTRACT. The deformation of a two-dimensional membrane due to the flow of an incompressible fluid around it is considered. The membrane is assumed to be flexible and inextensible. The problem is related to the flow of air past parachutes. Numerical solutions are obtained by discretization of the model nonlinear integrodifferential equation describing the flow. The numerical results are shown to be in reasonable agreement with experiments. In addition some aspects of the effect of porosity are discussed.

I. Introduction. We consider the deformation of a two-dimensional membrane due to the steady potential flow of an incompressible inviscid fluid around it (see Figure 1). We assume that the membrane is flexible, inextensible and characterized by a constant tension T. This flow configuration is relevant to the flow of air past a parachute. We shall first neglect the porosity of the cloth. Some aspects of porosity will be discussed in the last section of the paper.

We approximate the wake behind the membrane by assuming that the pressure in it is equal to a constant P_b.

As we shall see, the shape of the membrane is characterized by the dimensionless parameters

$$C_P = \frac{P_b - P_\infty}{\frac{1}{2}\rho U^2} \tag{1.1}$$

$$C_T = \frac{T}{\frac{1}{2}\rho U^2 C} \tag{1.2}$$

$$C_D = \frac{D}{\frac{1}{2}\rho U^2 C} \tag{1.3}$$

$$s = \frac{\ell - C}{C} . \tag{1.4}$$

Here C_P is the pressure coefficient, C_T the tension coefficient, C_D the drag coefficient, P_∞ the pressure far upstream, ρ the density of air, C

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the National Science Foundation under Grant No. MCS-800-1960.

the distance between the edges of the membrane, U the velocity far upstream, D the drag per unit span of the membrane and ℓ the length of the membrane.

This problem was previously considered by Newman and Low [1]. They presented experimental measurements of the drag coefficient and of the pressure coefficient for various values of s. In addition they performed some numerical computations based on the model of Parkinson and Jandali [2]. Their numerical results were found to be in satisfactory agreement with the experimental data.

In the present paper we compute accurate solutions by using the open-wake model. This model was introduced by Joukowsky, rediscovered by Roshko and extended by Wu [3]. Our numerical results are found to be in good agreement with the experimental data of Newman and Low [1].

The problem is formulated in Section 2 and the numerical results are discussed in Section 3. Some aspects of porosity are discussed in Section 4.

II. Formulation.

Let us consider the steady two-dimensional potential flow of an inviscid incompressible fluid past a membrane (see Figure 1). We use the open-wake model to describe the wake behind the membrane. In this model the portions CD and C'D' of the boundary of the wake are described by the free streamline theory. Downstream of the points D and D' the boundary of the wake consists of two horizontal straight lines. The position of the points D and D' is to be found as part of the solution.

It is convenient to introduce dimensionless variables by using U as the unit velocity and C as the unit length. We introduce the potential function ϕ and the stream function ψ. Without loss of generality we choose $\phi = 0$ at the point B and $\psi = 0$ on the membrane and on the surface of the wake. The flow configuration in the complex potential plane $f = \phi + i\psi$ is illustrated in Figure 2. We denote by b and d the values of ϕ at C and D. These values will be determined as part of the solution.

We denote the complex velocity by $u - iv$ and we define the function $\tau - i\theta$ by the relation

$$u - iv = e^{\tau - i\theta} . \qquad (2.1)$$

We shall seek $\tau - i\theta$ as an analytic function of $f = \phi + i\psi$ in the half plane $\psi < 0$.

The Bernoulli equation yields

$$\frac{1}{2} q^2 + \frac{P}{\rho} = \frac{1}{2} U^2 + \frac{P_\infty}{\rho} , \quad \psi < 0 . \qquad (2.2)$$

Here q is the magnitude of the velocity and P the pressure. In dimensionless variables (2.2) becomes

$$e^{2\tau} + \frac{P}{\frac{1}{2}\rho U^2} = 1 + \frac{P_\infty}{\frac{1}{2}\rho U^2} , \quad \psi < 0 . \qquad (2.3)$$

On the portion CD of the surface of the cavity we require $P = P_b$. Therefore (2.3) yields

$$e^{2\tau} = 1 - C_p, \quad b < \phi < d, \ \psi = 0 \ . \tag{2.4}$$

Here C_p is defined by (1.1).

On the membrane surface the Bernoulli equation (2.3) and the pressure jump due to the tension T yield

$$e^{2\tau} + C_T e^{\tau} \frac{\partial\theta}{\partial\phi} = 1 - C_p, \quad 0 < \phi < b, \ \psi = 0 \ . \tag{2.5}$$

Here C_T is defined by (1.2).

On the portion DE of the boundary of the wake we impose the condition

$$\theta = 0, \quad d < \phi < \infty, \quad \psi = 0 \ . \tag{2.6}$$

Finally, the symmetry of the problem yields

$$\theta = 0, \quad -\infty < \phi < 0, \ \psi = 0 \ . \tag{2.7}$$

At infinity we require the velocity to be 1 in the x-direction. Therefore, $\tau - i\theta$ must tend to zero at infinity. The function $\tau - i\theta$ is analytic in the half plane $\psi \leq 0$ and vanishes at infinity. Therefore on $\psi = 0$ its real part is the Hilbert transform of its imaginary part. Using (2.6) and (2.7) the Hilbert transform yields

$$\tau(\phi) = \frac{1}{\pi} \fint_0^d \frac{\theta(\phi')}{\phi' - \phi} \, d\phi' \ . \tag{2.8}$$

The integral in (2.8) is of Cauchy principal value form. The functions $\tau(\phi)$ and $\theta(\phi)$ in (2.8) denote the values of τ and θ on $\psi = 0$.

Relation (2.1) yields the indentity

$$\frac{\partial x}{\partial\phi} + i \frac{\partial y}{\partial\phi} = e^{-\tau+i\theta} \ . \tag{2.9}$$

Taking the imaginary part of (2.9) and integrating from 0 to b we obtain

$$\int_0^b e^{-\tau} \sin\theta \, d\phi + \frac{1}{2} = 0 \ . \tag{2.10}$$

Relation (2.10) defines the unit length as C.

Finally, the length of the membrane and the drag coefficient are given by

$$s = \frac{\ell - C}{C} = 2 \int_0^b e^{-\tau} d\phi - 1 \ , \tag{2.11}$$

$$C_D = (1 - C_P) - 2 \int_0^b e^{2\tau} y_\phi d\phi \ . \tag{2.12}$$

For given values of C_P and s, relations (2.4), (2.5), (2.8), (2.10), (2.11) and (2.12) define a nonlinear system of integrodifferential equations for the unknowns $\theta(\phi)$, $\tau(\phi)$, b, d, C_D and C_T. This system was discretized and the resultant algebraic equations were solved by Newton's method. The details of the numerical procedure follow closely the work of Vanden-Broeck [4]. Therefore they will not be repeated here.

III. Discussion of the results. Roshko [5] derived an exact solution for the open-wake model past a flat plate. In particular he calculated the following expression for C_D

$$C_D = \frac{\pi}{2} \{W^3(1 + W^2)^{-1} + W^2(1 - W^2)^{-1}[\frac{\pi}{2} - (1 + W^2)\tan^{-1}W]\}^{-1} \ . \tag{3.1}$$

Here

$$W = (1 - C_P)^{-1/2} \ .$$

In order to check the accuracy of our numerical procedure, we run our program with $C_P = -1.34$ and $s = 0$ for 20, 40, 100 and 150 mesh points. The respective values of C_D were 2.104, 2.095, 2.092 and 2.091. This sequence of values is in good agreement with the exact value 2.091 predicted by (3.1). This constitutes an important check on the validity of our numerical approach.

The experimental results of Newman and Low [1] suggest that C_P is a linear function of s for $0 < s < 0.6$. We found that the experimental values presented in Figure 13 of their paper could be fitted by the straight line

$$C_P = -1.3 - 0.6s \ . \tag{3.2}$$

The numerical scheme was run for various values of s and C_P satisfying (3.2). Some of the results are presented in Table 1.

s	C_P	C_D	C_T
0.02	-1.31	2.16	3.34
0.1	-1.36	2.28	1.71
0.2	-1.42	2.38	1.41
0.3	-1.48	2.45	1.31
0.4	-1.54	2.52	1.29
0.6	-1.66	2.66	1.33

Table 1

The values of C_D and C_T given in Table 1 agree within ten percent with the experimental data presented by Newman and Low in Figures 8 and 18 of their paper. Therefore the open wake is adequate for the prediction of C_D and C_T.

IV. The effect of porosity. Cumberbatch [6] extended the open-wake model to the flow past a porous flat plate. He assumed a Darcian flow through the mesh and obtained a solution in closed form. In addition he showed that his solution is in fair agreement with the experimental data of Taylor and Davies [7].

Numerical solutions for the flow past a flexible and porous membrane can be obtained by an appropriate generalization of the procedure described in Section 2. When porosity is included the stream function ψ is no longer a constant along the membrane. The basic idea is to write $\psi = F(\phi)$ along the membrane and to determine $F(\phi)$ by imposing Darcy's law across the mesh. Preliminary computations with $s = 0$, gave results in agreement with Cumberbatch's [6] solution. Work is continuing on this aspect of the problem.

REFERENCES

1. Newman, B. G. and Low, H. T. 1981. Aeronautical Quarterly 32, 243.

2. Parkinson, G. V. and Jandeli, T. 1970. J. Fluid Mech. 40, 577.

3. Wu, T. Y. 1962. J. Fluid Mech. 13, 161.

4. Vanden-Broeck, J.-M. 1982. Phys. Fluids 25, 420.

5. Roshko, A. 1954a. NACA TN 3168.

6. Cumberbatch, E. 1982. Q. J. Mech. Appl. Math. 35, 335.

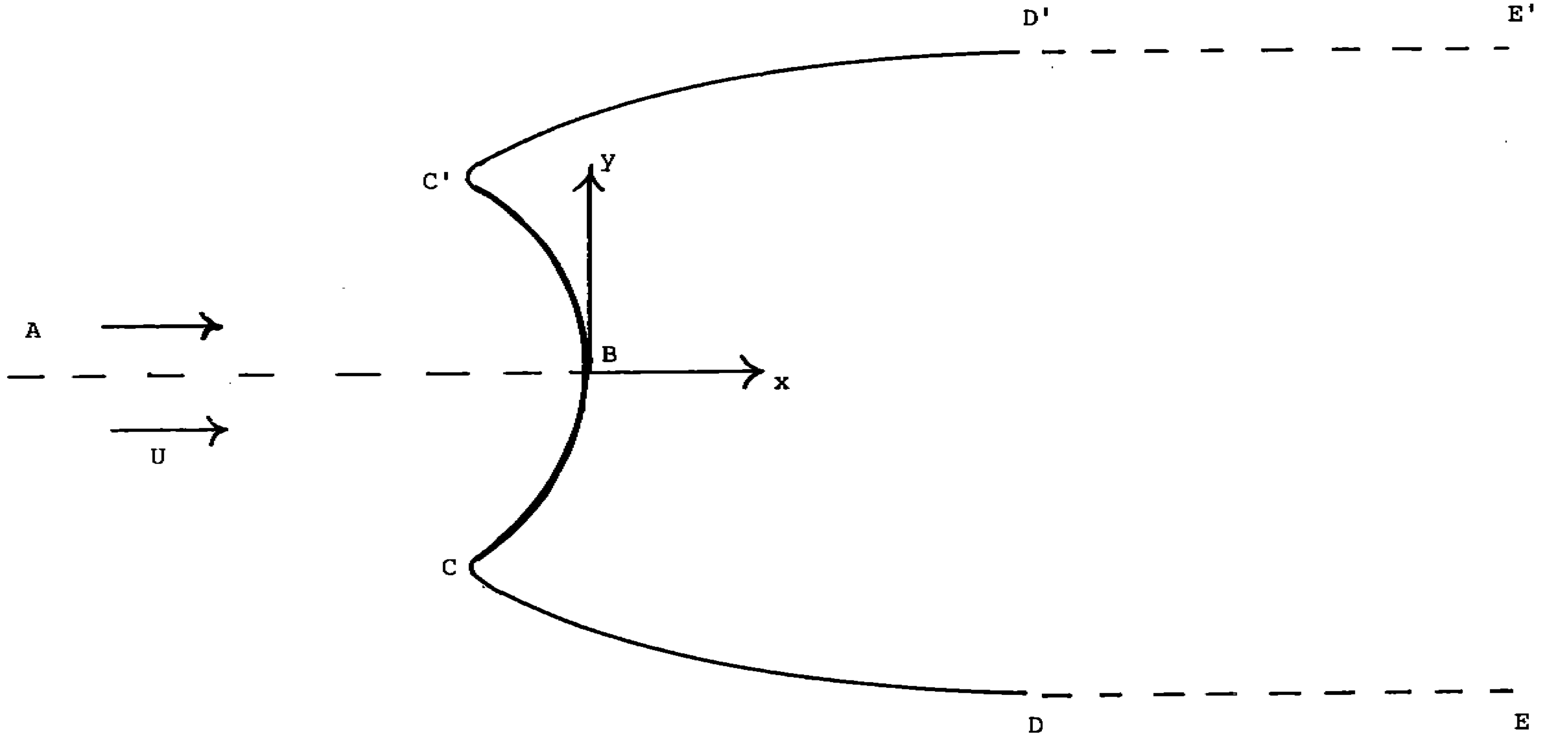


Figure 1: Sketch of the flow

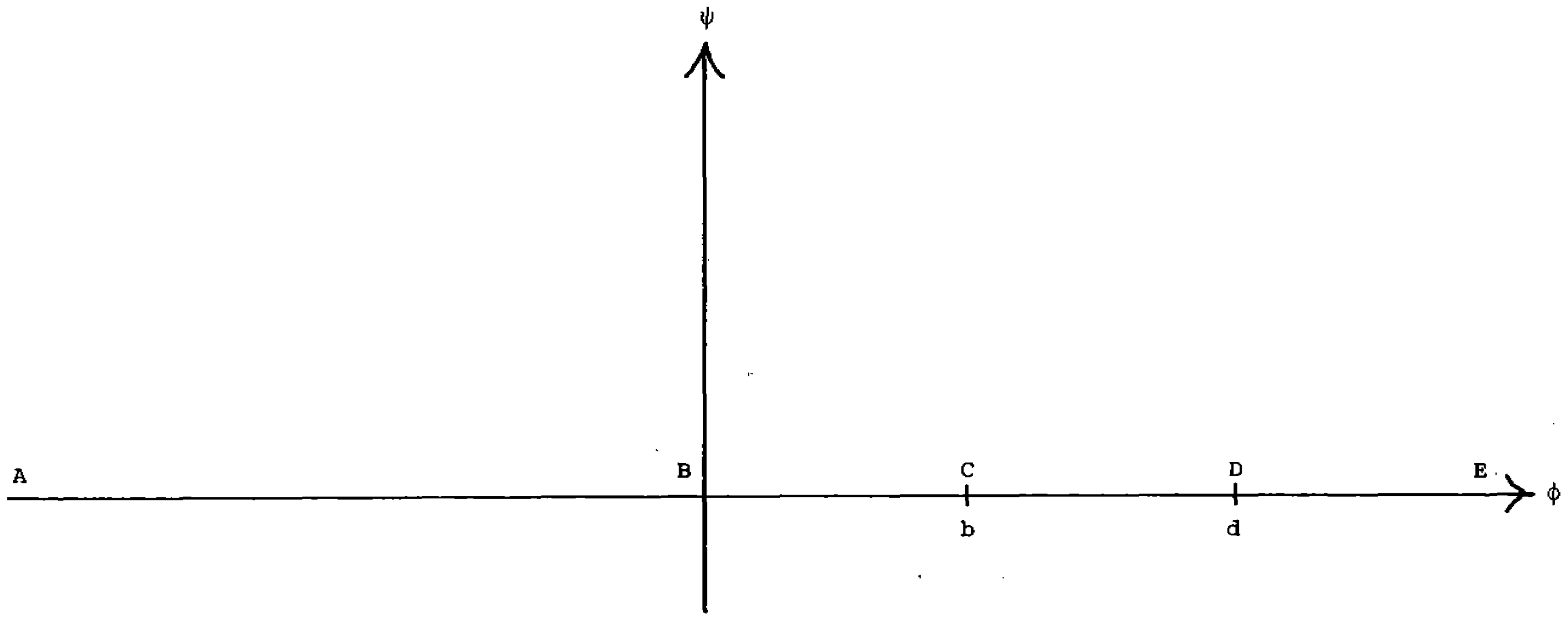


Figure 2: The complex potential plane

VEHICLE DYNAMIC ANALYSIS WITH FLEXIBLE COMPONENTS*

Sang Sup Kim, Ahmed A. Shabana, and Edward J. Haug
Center for Computer Aided Design
College of Engineering
The University of Iowa
Iowa City, Iowa 52242

ABSTRACT. A method is presented for nonlinear, transient dynamic analysis of vehicle systems that are composed of interconnected rigid and flexible bodies. The finite element method is used to characterize deformation of each elastic body and a component mode technique is employed to reduce the number of elastic generalized coordinates. Equations of motion and constraints of the coupled system are formulated in terms of a minimal set of modal and reference generalized coordinates. A Lagrange multiplier technique is used to account for kinematic constraints between bodies and a generalized coordinate partitioning technique is employed to eliminate dependent coordinates. The method is applied to a planar truck model with a flexible chassis and nonlinear suspension components. Simulation results for transient dynamic response as the vehicle traverses a bump, including the effect of bump-stops, and random terrain show that flexibility of the chassis can be routinely accounted for and predicts significant effects on vibratory motion of the vehicle. Compared with a rigid body model, flexibility of the chassis increases peak acceleration of the chassis and induces high frequency vertical acceleration in the range of human resonance, which deteriorates ride quality of off-road vehicles.

1. INTRODUCTION.

Modern, lightweight, off-road vehicle systems, operating over rough terrain, have placed increasingly higher demands on the technology required to accurately model and predict dynamic response of a vehicle system. In order to predict dynamic performance of a vehicle, it is necessary to consider nonlinear suspension kinematics and forces, coupled with elastic deformation of the vehicle chassis. Accurate description of vehicle dynamics; e.g., ride comfort and precision of armament subsystems, requires a high resolution mathematical model that accounts for flexibility effects and their coupling with geometrical and suspension force nonlinearities. This is mainly due to the large number of degrees-of-freedom required to model vehicle components and the high degree of geometrical nonlinearity associated with gross motion of suspension components and force-displacement nonlinearity associated with suspension bump-stops. When flexibility is considered, the problem becomes even more difficult, because of the increasing dimensionality and high frequencies of natural vibration.

Some investigators [1-2] have considered flexibility of vehicle components. Their method of analysis is based on a linear theory that has been employed to analyze mechanisms with flexible members [3-5]. In this

*Research Supported by Project No. DAAG29-82-K-0086 U.S. Army Research Office.

analysis, elastic deformation is assumed to have no significant effect on gross motion. Gross motion is first determined, using rigid body analysis and the resulting inertia and reaction forces are introduced in elastic analysis of the components. The total motion of the elastic member is then obtained by superposition of small deformation on gross body motion. There is, however, an increasing demand to produce lighter weight vehicular components that operate at higher speeds. Linear theory assumptions are no longer accurate enough to represent system dynamics, since flexibility effects can significantly affect motion at the driver's station.

Sunada and Dubowsky [6] recently presented a method for the dynamic analysis of flexible mechanisms that couples flexible degrees of freedom with a geometrically nonlinear set of equations of motion. Existing finite element structural programs are combined with a 4×4 matrix dynamic analysis technique. The method has been applied to analyze spatial mechanisms and robotic manipulators. The capability of this method to treat applications such as vehicle systems and space structures without substantial ad-hoc formulation, is not clear. Further, this method neglects rotary inertia of mass that is lumped at individual grid points, in order to avoid the difficulty of using a consistent mass approach to represent inertia coupling between the rigid body motion and the elastic deformation.

Shabana and Wehage [7-8] presented a method for dynamic analysis of large scale inertia-variant flexible systems with coupled reference and elastic deformation. Each flexible body is represented by two sets of generalized coordinates. The first set defines the location and orientation of a body-fixed coordinate system that is rigidly attached to a point on the flexible body. Second, elastic generalized coordinates characterize small deformation relative to the body-fixed system. This set of coordinates is introduced using the finite element method of structural analysis. Modal analysis is employed to reduce the number of elastic degrees of freedom, hence reducing problem dimensionality to manageable extent.

The purpose of this paper is to adapt the automated analysis method of Refs 7-8 for coupled dynamic analysis of planar vehicle systems that are composed of rigid and flexible bodies. As a numerical example, a cross country truck is considered in which the chassis is flexible. This investigation is mainly concerned with analysis of the effect of chassis flexibility on dynamic response of the vehicle, over a single bump and random terrain. A rather simplified tire model is used in this study. Extension of the formulation presented and illustrated here to include a more realistic tire model and three dimensional structural vibration [9] is theoretically simple but involves more detailed calculations than are presented here.

To achieve the above goals, the DADS computer program is used to automatically generate equations of motion, using a Lagrangian formulation. The system equations of motion are solved numerically using a direct integration method and advantage is taken of sparsity of the matrices arising in the formulation. As is shown, this automated formulation is general and conserves manpower that would be required in ad-hoc model formulation and analysis.

2. VEHICLE AND ROAD SURFACE MODELS.

2.1 Vehicle Model. The vehicle used in this investigation is a 5 ton, cross country truck [10] with rigid axles and a Watt mechanism independent suspension. Figure 1 shows the general configuration of the truck. Figure 2 illustrates the wheel suspension, with two suspension support arms beneath the axle and two arms above. Vertical forces are supported by coil springs of progressive stiffness. Figure 3 is an overall view of the frame, which consists of two sidemembers (closed box griders) and several tubular crossmembers that are mounted in holes in the sidemembers and welded to their inner and outer sides. Vehicle parameters used in this analysis are given in Table 1.

Table 1 Vehicle Parameters

Parameter		Value
Gross Vehicle Mass (including payload)		14,400 kg
Vehicle Sprung Mass		11,950 kg
Vehicle Unsprung Mass		2,450 kg
Front axle		1185.0 kg
Rear axle		1185.0 kg
Long trailing arms		44.8 kg
Short trailing arms		35.2 kg
Pitch moment of inertia of sprung mass (about C.M.)		58300.0 kg-m^2
Front and Rear Suspension		
Spring rate (per spring)		6.91x10^5 N/m
Damping rate (per shock absorber)		
Compression		5480.0 N.sec/m
Rebound		17575.0 N.sec/m
Wheel Travel (unloaded)		
Jounce		0.15 m
Tire	Quadratic Spring Constant (per tire)	5.649x10^7 N/m^2
	Damping rate (per tire)	4625.0 N.sec/m
Tire Radius		0.6 m
Vertical Natural Frequency of Sprung Mass		1.98 Hz
Sprung Pitch Natural Frequency		1.95 Hz

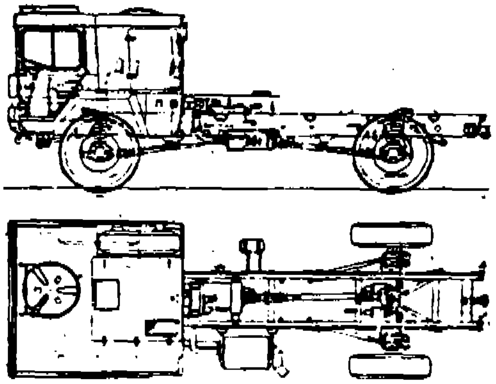


Fig. 1 5 Ton 4x4 Cross Country Truck

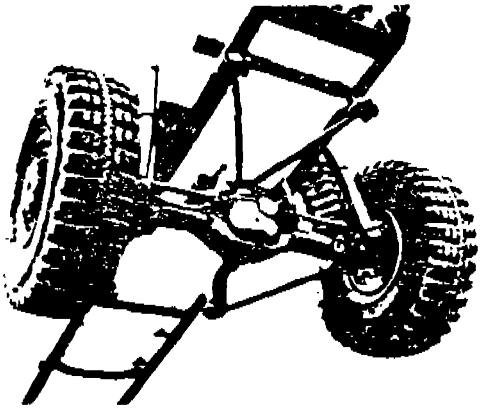


Fig. 2 Wheel Suspension

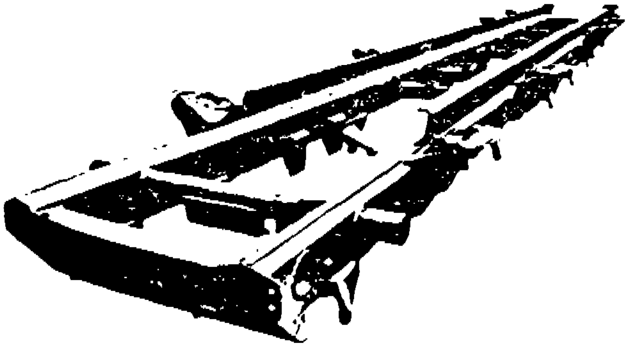


Fig. 3 Main Frame

A simplified, planar rigid body truck model is shown in Fig. 4. Bodies 1 and 2 are the front and rear axle-wheel assemblies, respectively. Each axle-wheel mass is assumed to be concentrated at the wheel center. Body 3 is the chassis of the truck. The mass of the chassis includes masses of the payload and engine. Body-fixed coordinate axes are located at the centroid of each body. Bodies 4, 5, 6 and 7 are the trailing arms that connect the chassis and wheel assemblies by revolute joints. The function of these trailing arms is to provide kinematic control of the axle position and to absorb driving and braking torques acting on the wheels. Therefore, they are modeled as a Watt's mechanism, which gives very small rotation to the axle during vertical displacement. Rigid body data and the initial location and orientation of the body-fixed coordinate systems of each body, with respect to the inertial reference frame, are given in Table 2.

The suspension springs and dampers and the tires are modeled by springs and dampers, as shown in Fig. 5. Spring characteristics of the tires are taken here as quadratic functions of displacement. A simple point contact tire model is used to simulate tire forces that occur due to motion of the wheel relative to the road surface. Fore-and-aft force components are neglected, assuming the tire force is always vertical. The tire is free to leave the ground, to simulate wheel hop. Nonlinear spring and damping characteristics of suspension elements are given in Fig. 6. The high stiffness of the suspension spring in compression, when the spring deflection is greater than 0.15 m, simulates the bump-stop in the suspension system.

A rigid body vehicle model of this vehicle, with five degrees-of-freedom (5-DOF), is also formulated to allow evaluation of the effects of chassis flexibility.

Table 2 Rigid Body Data and Initial Positions

Body No.	Mass (kg)	Moment of Inertia about C.G. (Kg-m^2)	Initial Body Coordinates X(m)	y(m)	ϕ(rad)
1	1185	13.33	1.500	0.575	0.0
2	1185	13.33	5.850	0.575	0.0
3	11950	58300	3.675	0.975	0.0
4	22.4	2.38	2.116	0.725	0.464
5	22.4	2.38	5.234	0.725	-0.464
6	17.6	1.14	1.066	0.909	-0.154
7	17.6	1.14	6.284	0.909	0.154

2.2 Road Surface Models. The dynamic response of a vehicle depends strongly on the vertical displacement history of the wheels on the road surface. In this investigation, two roadway models are used, as shown in Fig. 7. Figure 7(a) represents a simulated obstacle with 0.2m height and 0.4m

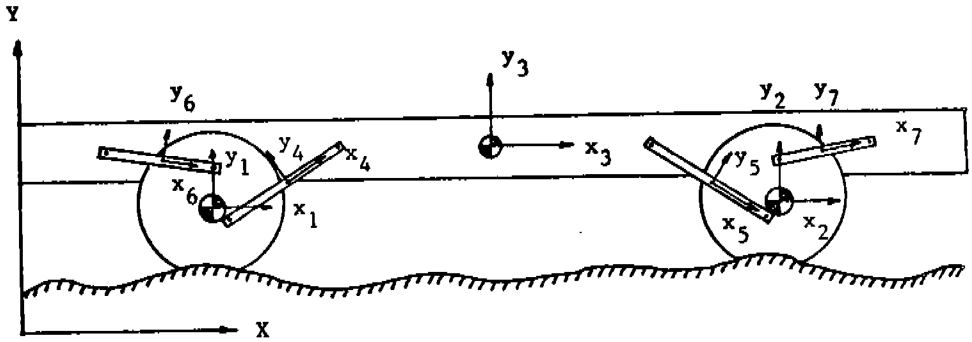


Fig. 4 Rigid Body Truck Model

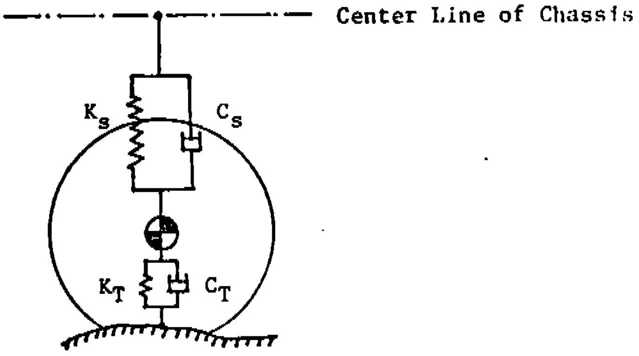
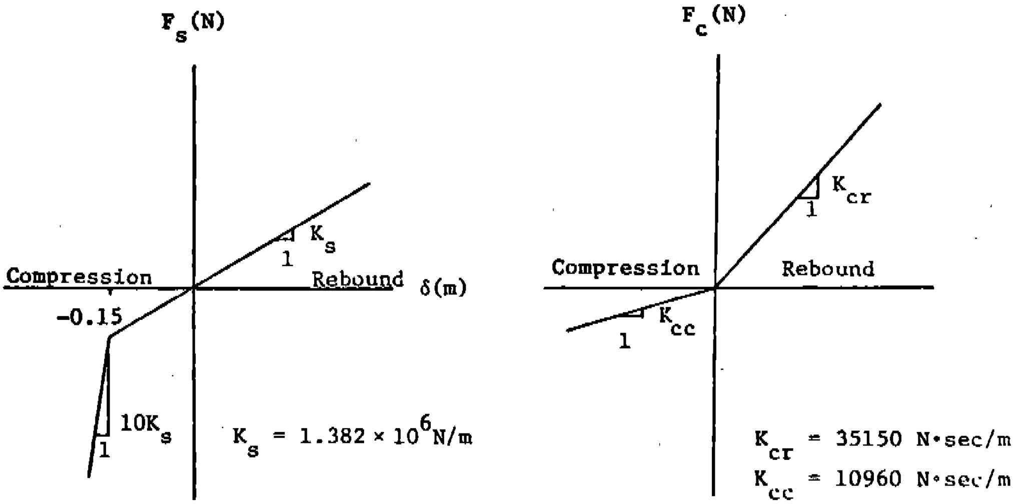


Fig. 5 Suspension and Tire Model



(a) Spring Characteristics

(b) Damping Characteristics

Fig. 6 Suspension Characteristics

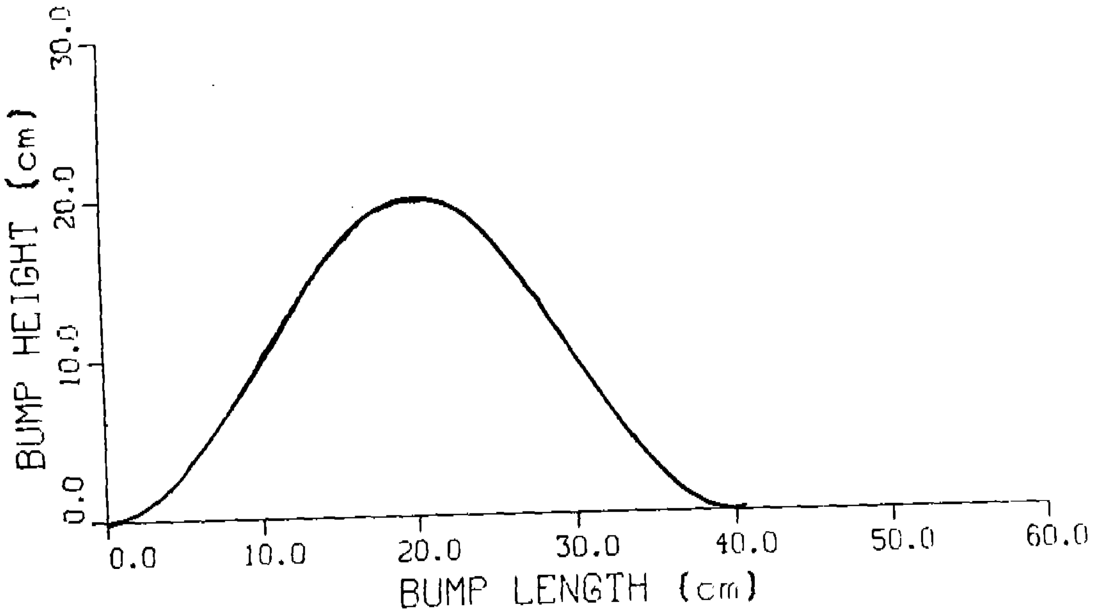


Fig. 7(a) Single Bump Road Profile

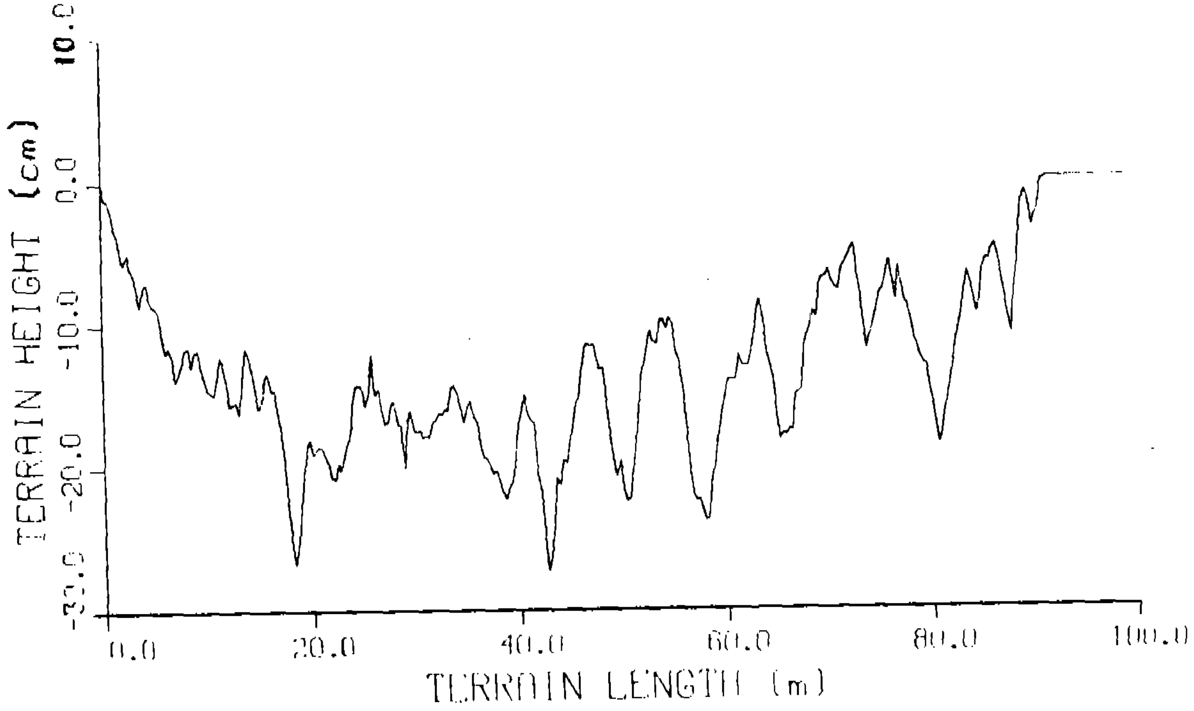


Fig. 7(b) Random Terrain Profile

width, which is used to simulate shock response of the vehicle over a single bump. Figure 7(b) is a terrain profile with a RMS roughness of 2.64cm (1.04 in) and a length of 91.44m (300 ft).

3. ANALYTICAL APPROACH. The analysis method employed in this investigation is similar to the method used in Refs. 7-8 to analyze mechanical systems with interconnected rigid and flexible bodies. In this method, the chassis of the vehicle is considered as a deformable substructure. Two sets of generalized coordinates are employed to describe the flexible body configuration. First, reference generalized coordinates define the location and orientation of a body-fixed coordinate system on each body. Second, a set of elastic coordinates define small deformation of each body, relative to its body-fixed coordinate system. This set is introduced using the finite element method.

Kinetic and strain energy expressions are developed for the individual elements. The kinetic and strain energy of each body are obtained by summing energies of its elements. Constraints between different elements of a body are expressed in a Boolean form and constraints between different bodies are introduced using a Lagrange multiplier technique. The generalized coordinate partitioning method [11] and a component mode structural analysis technique are employed to describe the system equations of motion, with a minimal set of independent generalized coordinates [7,8]. The method of Refs. 7 and 8 is summarized here, for completeness.

3.1 Energy Expressions. Figure 8 shows a typical element j of a two dimensional planar flexible body i. Let the x-y coordinate system represent an inertial reference frame and the x^i-y^i system represent a coordinate system that is rigidly attached to body i.

The location of an arbitrary infinitesimal volume at point p^{ij} on element j can be defined as

$$\underline{R}_p^{ij} = \underline{R}^i + A^i \underline{r}^{ij} \tag{1}$$

where $\underline{R}^i = [x^i, y^i]^T$ is the vector of translational coordinates of the origin of the coordinate system of body i with respect to the x-y system,

$$A^i = \begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} \tag{2}$$

is the transformation matrix from x^i-y^i to x-y coordinate systems, and $\underline{r}^{ij}$ is the position vector of p^{ij} with respect to the x^i-y^i system, defined as

$$\underline{r}^{ij} = \underline{r}_0^{ij} + \underline{W}^{ij} \tag{3}$$

where $\underline{r}_0^{ij}$ is the position vector of p^{ij} in the undeformed state and $\underline{W}^{ij}$ is the elastic displacement vector in the body fixed coordinate system. Let an x^{ij}-y^{ij} coordinate system be attached to the left end of element j. Using a shape function, $\underline{r}^{ij}$ can be expressed in terms of nodal coordinates $\underline{e}_k^{ij}$ (k = 1,2...6), which represent nodal coordinates and slopes of reference lines at nodes, relative to the x^i-y^i system,

$$\underline{r}^{ij} = N^{ij}\underline{e}^{ij} \tag{4}$$

where N^{ij} is the element shape function.

From Eq. 1, the position vector $\underline{R}_p^{ij}$ can be expressed [7-8], in terms of reference coordinates (x^i, y^i, θ^i) and nodal coordinates $(\underline{e}^{ij})$, as

$$\underline{R}_p^{ij} = \underline{R}^i + A^i N^{ij}\underline{e}^{ij} \tag{5}$$

Differentiating Eq. 5 with respect to time gives

$$\dot{\underline{R}}_p^{ij} = \dot{\underline{R}}^i + \dot{A}^i N^{ij}\underline{e}^{ij} + A^i N^{ij}\dot{\underline{e}}^{ij} \tag{6}$$

where

$$\begin{aligned}\dot{A}^i &= \dot{\theta}^i \begin{bmatrix} -\sin\theta^i & -\cos\theta^i \\ \cos\theta^i & -\sin\theta^i \end{bmatrix} \\ &= \dot{\theta}^i A^{i'}\end{aligned} \tag{7}$$

Substituting Eq. 7 into Eq. 6 and writing the result in partitioned form yields

$$\dot{\underline{R}}_p^{ij} = \begin{bmatrix} I & A^{i'}N^{ij}\underline{e}^{ij} & A^i N^{ij} \end{bmatrix} \begin{Bmatrix} \dot{\underline{R}}^i \\ \dot{\theta}^i \\ \dot{\underline{e}}^{ij} \end{Bmatrix} \tag{8}$$

The kinetic energy expression for element ij is given by

$$\begin{aligned}T^{ij} &= \frac{1}{2}\int_{V^{ij}} \rho^{ij}\, \dot{\underline{R}}_p^{ij^T}\, \dot{\underline{R}}_p^{ij}\, dV^{ij} \\ &= \frac{1}{2}\, \dot{\underline{q}}^{ij^T} M^{ij}\, \dot{\underline{q}}^{ij}\end{aligned} \tag{9}$$

where V^{ij} is the element volume, ρ^{ij} is density of the element material,

$$\underline{q}^{ij} = [\ \underline{R}^{i^T}\ \theta^i\ \underline{e}^{ij^T}\ ]^T \tag{10}$$

are the generalized coordinates of element ij and M^{ij} is the element mass matrix [7-8]. The vector e^{ij} can be written as

$$\underline{e}^{ij} = \underline{e}_0^{ij} + \underline{q}_f^{ij} \tag{11}$$

where $\underline{e}_0^{ij}$ is the vector of nodal coordinates in the undeformed state and $\underline{q}_f^{ij}$ is the vector of deformations at the nodes, defined with respect to the body-fixed coordinate system.

The total kinetic energy of body i is given by

$$\begin{aligned} T^i &= \sum_{j=I}^{n^i} T^{ij} \\ &= \frac{1}{2}\dot{\underline{q}}^{i^T} M^i \dot{\underline{q}}^i \end{aligned} \tag{12}$$

where $\underline{q}^i = [\ \underline{R}^{i^T}\ \theta^i\ \ \underline{q}_f^{i^T}\ ] = [\ \underline{q}_r^{i^T}\ \ \ \underline{q}_f^{i^T}\ ]$, $\underline{q}_r^i$, and $\underline{q}_f^i$ represent, respectively, reference and elastic coordinates of body i. The strain energy of body i can also be expressed in compact form as [7-8]

$$U^i = \frac{1}{2}\underline{q}^{i^T} K^i \underline{q}^i \tag{13}$$

where K^i is the stiffness matrix of body i.

The virtual work of external forces acting on body i can be written as

$$\delta W^i = \underline{Q}^{i^T}\ \delta\underline{q}_i \tag{14}$$

where $\underline{Q}^i$ is the vector of generalized forces associated with the generalized coordinates of body i.

3.2 Equations of Constraint. When adjacent bodies are connected, nonlinear constraint equations are written between adjacent bodies and a Lagrange multiplier method is employed to adjoin these constraint equations to the equations of motion. These constraints permit the joining of elastic bodies, rigid bodies, or rigid and elastic bodies. Points of attachment on elastic bodies are at nodes of the finite element model. In general, equations of constraint can be written, in vector function form, as

$$\underline{\Phi}(\underline{q},t) = 0 \tag{15}$$

where $\Phi(q,t) = [\,\phi_1(q,t),\ldots,\phi_m(q,t)]^T$. This is a set of nonlinear algebraic equations, which can be used to describe constraints between vehicle components.

3.3 Equations of Motion. The composite vector of all system generalized coordinates is designated as $\underline{q} = [\underline{q}^{1T}, \underline{q}^{2T}. \cdots\, \underline{q}^{N^T}]^T$, where N is the total number of bodies (substructures) in the system. The constraint equations of Eq. 15 are assumed to be independent. Presuming that the constraints are workless, the variational form of the equations of motion [12] for body i, where subscript notation denotes differentiation with respect to a vector, is

$$\frac{d}{dt}\, T^i_{\dot{\underline{q}}^i} - T^i_{\underline{q}^i} + U^i_{\underline{q}^i} - \underline{Q}^{iT}\quad \delta\underline{q}^i = 0 \tag{16}$$

for all virtual displacements $\delta\underline{q}^i$ that are consistent with constraints of Eq. 15. It can be shown that introducing the vector $\underline{\lambda}^T\,\underline{\Phi}_{qi}$ into Eq. 16 allows the coefficients of $\delta\underline{q}^i$ to be set to zero [13]. Thus,

$$\frac{d}{dt}\, {T^i_{\dot{\underline{q}}^i}}^T - {T^i_{\underline{q}^i}}^T + {U^i_{\underline{q}^i}}^T - \underline{Q}^i + \underline{\Phi}^T_{\underline{q}^i}\underline{\lambda} = 0 \tag{17}$$

with T^i, U^i, and $\underline{Q}^i$ given by Eqs. 12, 13, and 14.

3.4 Generalized Coordinate Reduction. Efficient solution of the system equations of motion requires a transformation from the space of system nodal generalized coordinates to the space of system modal generalized coordinates, which has lower dimension. The method presented in Refs 7-8 is based on solving the eigenvalue problem for each substructure once. From Fourier analysis of the forcing functions, an initial estimate of the number of modes to be retained is made. During the simulation, additional eigenvectors are recalled or deleted, as required. For the purpose of determining eigenvalues and eigenvectors, if a substructure is assumed to vibrate freely about a reference configuration, Eq. 17 yields

$$\bar{M}^i_{ff}\,\ddot{\bar{\underline{P}}}^i_f + \bar{K}^i_{ff}\,\bar{\underline{P}}^i_f = 0 \tag{18}$$

Where $\bar{M}^i_{ff}$ and $\bar{K}^i_{ff}$ are the mass and stiffness matrices associated with the nodal generalized coordinates and $\bar{\underline{P}}^i_f$ is the vector of elastic coordinates after imposing the body-fixed coordinate conditions. The stiffness matrix $\bar{K}^i_{ff}$ is positive definite, because the reference coordinate system is fixed. Equation 18 yields a set of eigenvectors and a modal matrix. A coordinate transformation from the physical nodal coordinates to modal coordinates is defined by

$$\underline{\bar{P}}_f^i = B_2^i \, \underline{\chi}^i \tag{19}$$

where B_2^i is the modal matrix, consisting of the eigenvectors obtained from Eq. 18 and $\underline{\chi}^i$ is a vector containing the modal coordinates. Using Eq. 19, the reference and nodal generalized coordinates are written in terms of reference and modal coordinates. A substantial reduction in problem dimensionality can be achieved by considering only significant modes.

4. NUMERICAL RESULTS. In order to take the effect of flexibility of vehicle components on global vehicle motion into account, the chassis (Body 3) and long trailing arm (Body 4) are modeled as elastic bodies. The chassis and long trailing arm are divided into 12 and 2 finite beam elements, respectively, with reference coordinates located at their midpoints. The flexural rigidity of each flexible member is calculated using the cross-sectional area of the beam and its material properties. Lateral and axial deformation are considered.

The flexible components are initially treated as substructures that are fixed at their midpoints. Since each beam element has 6 degrees-of-freedom, the flexible chassis has a total of 36 elastic degrees-of-freedom and each flexible link has 6 elastic degrees-of-freedom. The eigenvalue problem is solved for each of these substructures. The lowest six natural frequencies of the flexible chassis are 6.11, 6.11, 38.31, 38.31, 83.36, and 83.36 Hz, where the first four modes are bending modes and the fifth and sixth modes are axial vibration modes. The lowest two natural frequencies of the flexible links are 88.92 and 88.92 Hz. One percent structural damping is considered for every chassis mode of vibration.

4.1 Vehicle Response over Single Bump. The vehicle travels over the single bump given in Fig. 7(a), with a vehicle speed of 3 m/sec (6.7 miles/hr). The simulation is carried out for two vehicle models, rigid and flexible chassis, to evaluate flexibility effects of the chassis on vehicle motion. To compare higher mode effects of the flexible chassis model, 2- and 4-mode solutions are obtained. Figure 9 shows the vertical displacement of the center of mass of the chassis, from its static equilibrium position. The figure shows significant peak differences between vertical displacement of rigid and flexible chassis models. Figure 9 also shows that the contribution of higher vibration modes to vertical displacement of the chassis is negligible.

The vertical acceleration at the center of mass of the chassis is given in Fig. 10, for each model. It shows that chassis flexibility results in increased peak acceleration and significantly higher frequency content during passage over the bump. The effect of structural damping on the vehicle response is shown in Fig. 11. In this figure the vertical acceleration of the chassis with and without damping are plotted. The damped response decays with time, while the undamped response has a sustained oscillation.

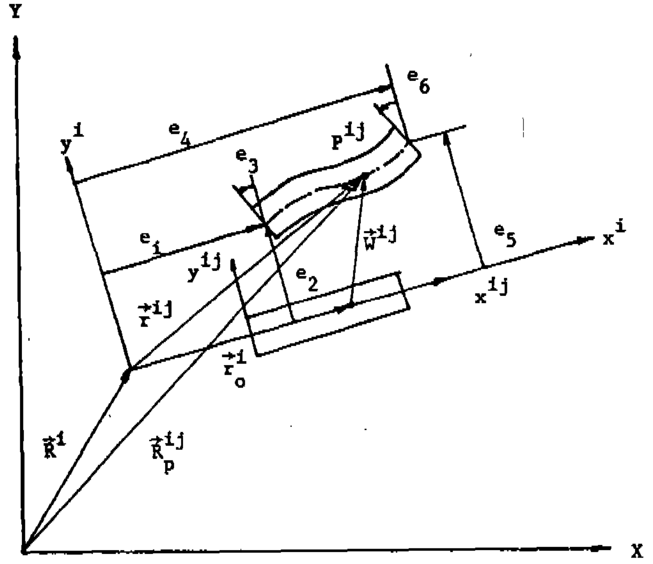


Fig. 8 Generalized Coordinates of a Beam Element

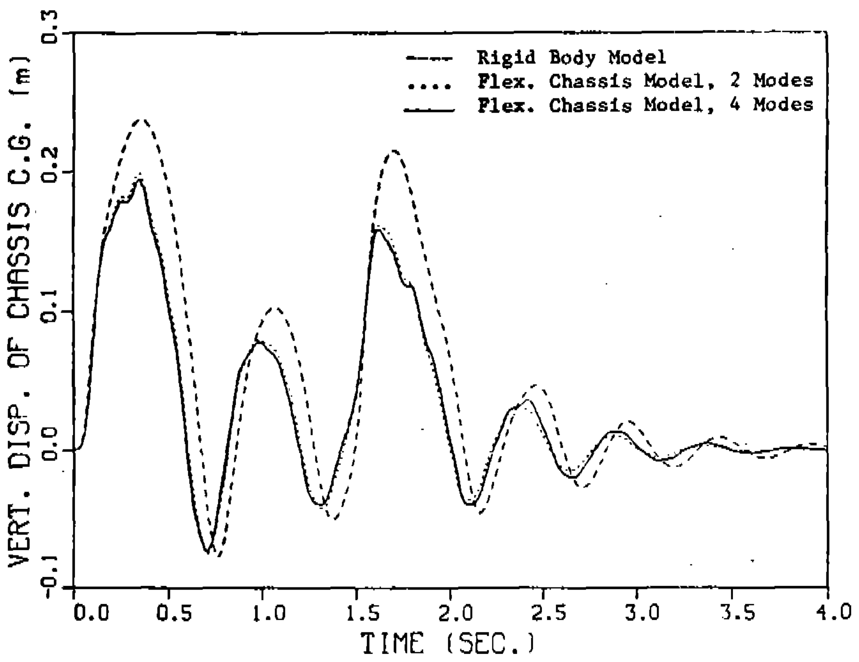


Fig. 9 Vertical Displacement of the Chassis C.G. over Single Bump

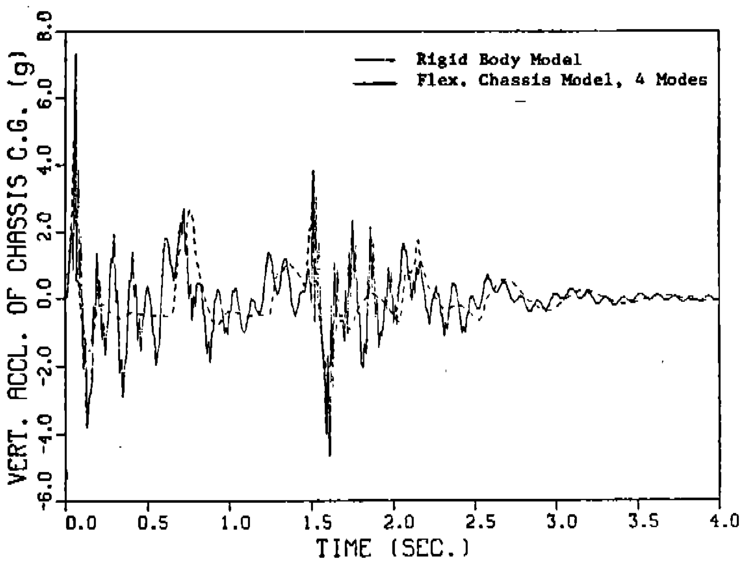


Fig. 10 Vertical Acceleration of the Chassis C.G. over Single Bump

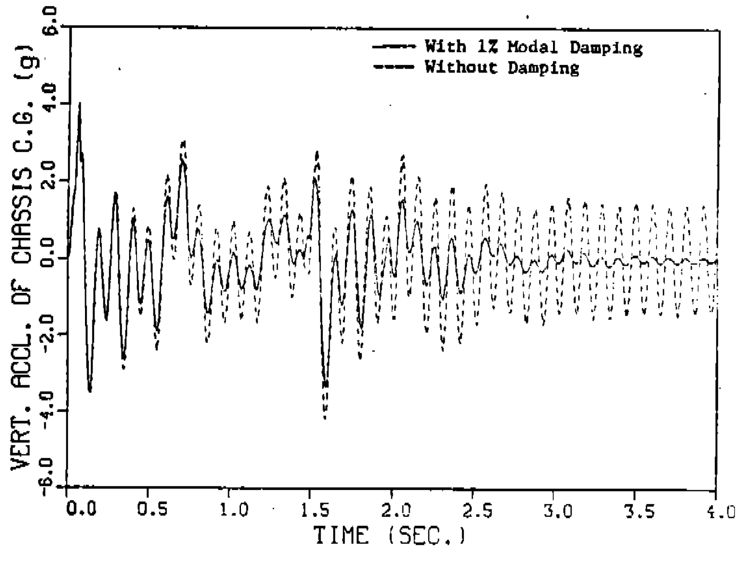


Fig. 11 Effect of Structural Damping on Vertical Acceleration of Chassis C.G. over Single Bump, Flex, Chassis Model, 2 Modes

Computation time for the rigid model simulation was 7.0 minutes on a PRIME 750 supermini computer. Computation times with the two and four mode flexible chassis models were 2.08 and 5.18 times the computation time of the rigid model. This shows that computational efficency can be obtained by using the smallest number of modes required to obtain reasonable accuracy.

4.2 Vehicle Response over Random Terrain. Simulation is carried out over the terrain given in Fig. 7(b) for the rigid model, the two and four mode flexible chassis models, the two mode flexible links model, and the two mode flexible chassis and links model. Vehicle velocity is 7.5 m/sec. (16.8 miles/h). Results are given in Figs. 12 to 14. Since link flexibility does not have significant effect on the global vehicle motion, results for the model with flexible links are not included.

Figure 12 shows vertical displacement at the center of mass of the chassis. No significant difference is observed between rigid and flexible models. No significant difference has been found between the two and four-mode solution of the flexible chassis models. It has also been found that there is no significant difference in vertical chassis displacement between the two mode flexible chassis and links model and the two mode flexible chassis model. It is concluded that flexibility effects of the stiff link (which has relatively high natural frequency) on the vehicle response may be neglected.

Figure 13 shows vertical acceleration at the center of mass of the chassis. Flexibility of the chassis results in a significant increase in the acceleration level at the center of the chassis and high frequency content near the resonant frequency of the human body. Vibration in the chassis may thus result in an unpleasant motion and deteriorate ride comfort of the vehicle. Suspension link flexibility does not have noticeable effects on the vertical acceleration of chassis.

Deflection of the front end of the chassis, with respect to its body-fixed coordinate system, is given in Fig. 14. Dynamic peak deflections for the two mode flexible chassis model is 22 times the static deflection of that model. The frequency of vibration of the front of the chassis is about 6 Hz, which is the fundamental natural frequency of the chassis.

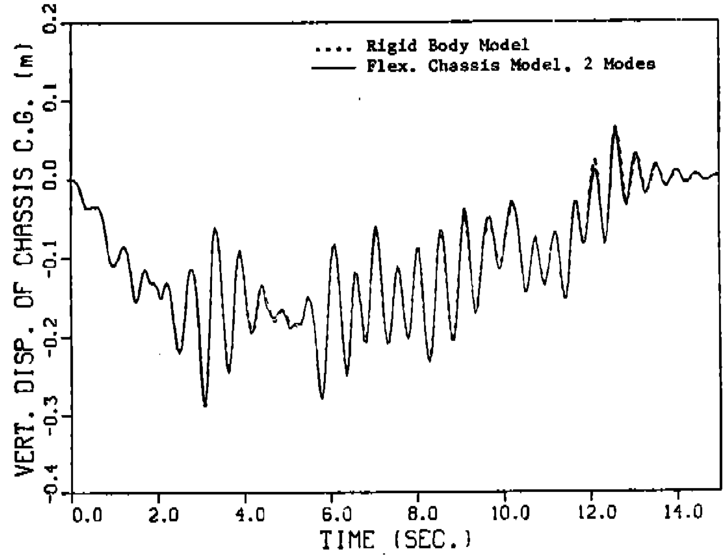
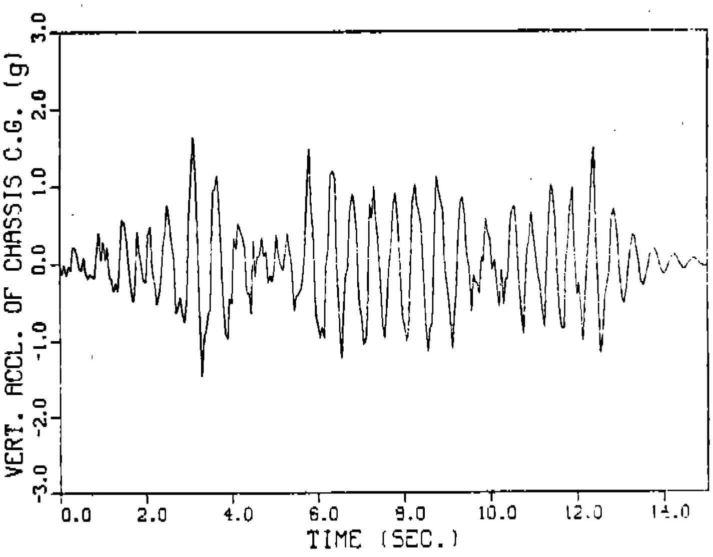


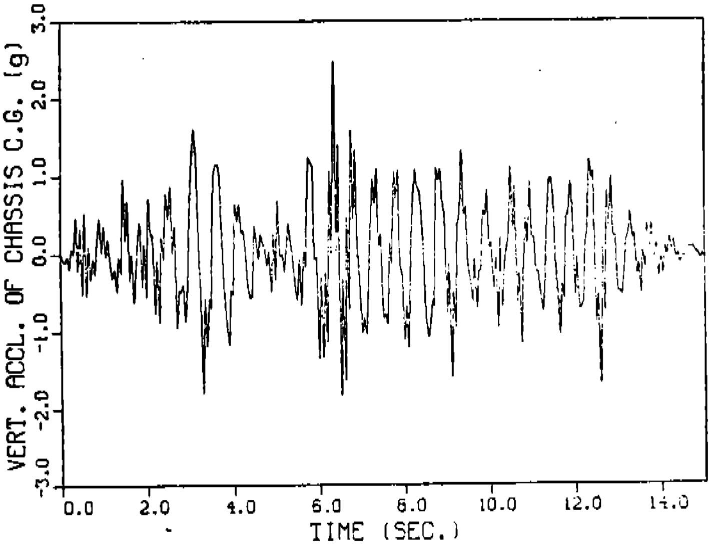
Fig. 12 Vertical Displacement of Chassis C.G. over Random Terrain



(a) Rigid Body Model

Fig. 13 Vertical Acceleration of Chassis C.G. over Random Terrain, 2 Modes

(Fig. 13 continued)



(b) Flex. Chassis Model

Fig. 13 Vertical Acceleration of Chassis C.G. over Random Terrain, 2 Modes

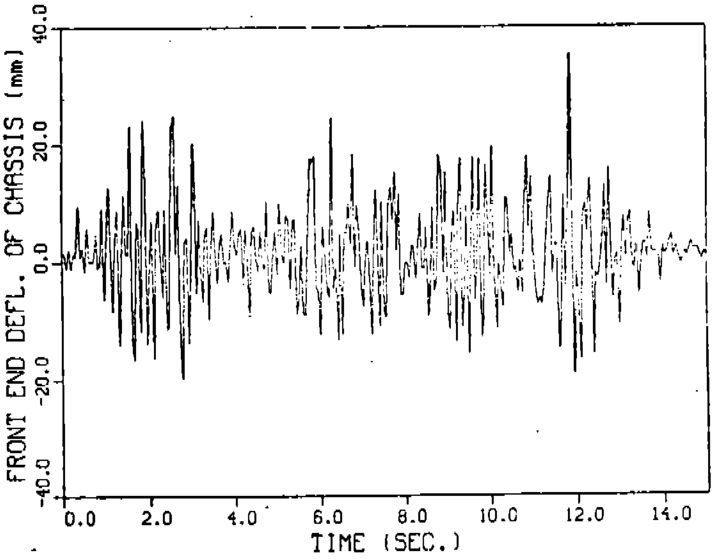


Fig. 14 Front End Deflection of Chassis over Random Terrain, Flex. Chassis Model, 2 Modes

References

1. Vail, C.F., "A Modal Synthesis Technique for Determining Properties for Structures for Mass and Stiffness Changes," SAE paper 740329, 1974.

2. Elmadany, M.M., Dokanish, M.A., and Allan, A.B., "Ride Dynamics of Articulated Vehicles - A Literatures Survey," Vehicle Systems Dynamics, Vol. 8, 1979, pp.289-316.

3. Winfrey, R.C., "Elastic Link Mechanism Dynamics," ASME Journal of Engineering for Industry, Feb. 1971, pp. 268-272.

4. Sadler, J.P., and Sandor, G.N., "A Lumped Parameter Approach to Vibration and Stress Analysis of Elastic Linkages," ASME Journal of Engineering for Industry, May 1973, pp. 549-557.

5. Erdman, A.C., Sandor, G.N., and Oakberg, R.G., "A General Method for Kineto-Elastodynamic Analysis and Synthesis," ASME Journal of Engineering for Industry, Nov. 1972, pp. 1193-1205

6. Sunada, W., and Dubowsky, S., "The Application of Finite Element Methods to the Dynamic Analysis of Flexible Spatial and Co-planar Linkage Systems," ASME Journal of Mechanical Design, July 1981, Vol. 103, pp. 643-651.

7. Shabana, A. and Wehage, R.A., "Variable Degree of Freedom Component Mode Analysis of Inertia-Variant Flexible Mechanical Systems," ASME paper No. 82-DET. 93, Journal of Mechanisms, Transmission and Automation in Design, to appear.

8. Shabana, A., and Wehage, R.A., "A Coordinate Reduction Technique for Dynamic Analysis of Spatial Substructures with Large Angular Rotations," Journal of Strucutral Mechanics, to appear, Dec. 1983.

9. Shabana, A.A. and Wehage R.A., "Spatial Transient Analysis of Inertia-Variant Flexible Mechanical Systems", Submitted to ASME Journal of Mechanical Design, Jan. 1983.

10. Klanner, R., "Cross-Country Truck for Fast Off-Highway Navigation, 6th International Conference of International Society of Terrain and Vehicle Systems," Vienna, Aug. 1978, pp. 22-25.

11. Wehage, R.A. and Haug, E.J., "Generalized Coordinate Partitioning for Dimension Reduction in Analysis of Constrained Dynamic Systems," ASME Journal of Mechanical Design, Vol. 104, 1982, pp. 247-255.

12. Goldstein, H., Classical Mechanics, Addison Wesley, 1980.

13. Haug, E.J. and Arora, J.S., Applied Optimal Design, Wiley 1979.

APPLICATION OF SYMBOLIC COMPUTATION TO THE ANALYSIS OF MECHANICAL SYSTEMS, INCLUDING ROBOT ARMS

M.A. Hussain, General Electric Company—CRD
B. Noble, Mathematics Research Center, University of Wisconsin

SUMMARY*

This paper illustrates the application of symbolic computation in connection with three aspects of mechanical systems:

1. The derivation of dynamical equations by Lagrangian methods.
2. The analysis and synthesis of kinematic mechanisms.
3. A robot manipulator arm.

INTRODUCTION

This paper illustrates the *potential* of symbolic computation in connection with the formulation and analysis of equations for dynamical systems, sensitivity analysis, linkages and mechanisms, and robot manipulator arms.

* This is a preliminary version of a paper to appear in the proceedings of NATO Conference on Mechanisms, University of Iowa, 1983.

We use MACSYMA (project MAC's SYmbolic MAnipulation system), a large-scale computer program for symbolic mathematical computation. MACSYMA can handle polynomial manipulation, simplification and substitution with symbolic expressions, symbolic solution of algebraic and differential equations, and matrix manipulation. Although we have found MACSYMA particularly convenient to use, other symbolic programs such as REDUCE and SMP could be used.

This report deals with three topics:

1. The derivation of dynamical equations by Lagrangian methods, including sensitivity analysis (Sections 1, 2, 3, 4, and 5).
2. The analysis and synthesis of kinematic mechanisms, including dual-number quaternions (Sections 7 and 8).
3. The direct and inverse problem involving robot manipulator arms (Sections 9 and 10).

In order to make the presentation clearer to the general reader who lacks specialized knowledge of symbolic manipulation, we explain the mathematical aspects in the main text (namely the kind of problem for which we feel symbolic computation is useful), and give the detailed MACSYMA programs in appendices.

In a certain sense the real "meat" of this paper is the detailed programs which appear in the appendices. The reader interested in symbolic manipulation should solve the problems outlined in the text using MACSYMA, or any other suitable program, with the appendices as a guide.

The objective of this paper is to encourage the use of symbolic manipulation in the analysis of mechanical systems. It is clear that the complexity of the problems being tackled is increasing to the point where symbolic manipulation must play an important role in their formulation and solution. In this paper we have simply picked out the tedious parts of well known methods and examples, and illustrated the ease of performing the manipulation using MACSYMA.

1. DESIGN OF A 5-DEGREES-OF-FREEDOM VEHICLE SUSPENSION

The objective of this example is to illustrate how MACSYMA handles Lagrange's equation of motion in the form:

$$\frac{d}{dt}\left[\frac{\partial T}{\partial \dot{q}_i}\right] - \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i} - Q_i = 0, \qquad i = 1, \cdots, n \tag{1}$$

where T and V are quadratic forms representing kinetic energy and potential energy, respectively, expressed in terms of generalized coordinates q_i. Q_i are nonconservative generalized forces. Consider the 5-degrees-of-freedom model of a vehicle suspension system shown in Figure 1 and dealt with in Haug and Arora [6] (pp. 25, 200):

$$\begin{aligned} T &= \frac{1}{2}\, m_1 \dot{Z}_1^2 + \frac{1}{2}\, m_2 \dot{Z}_2^2 + \frac{1}{2}\, I \dot{Z}_3^2 + \frac{1}{2}\, m_4 \dot{Z}_4^2 + \frac{1}{2}\, m_5 \dot{Z}_5^2 \\ V &= \frac{1}{2}\, K_1 d_1^2 + \frac{1}{2}\, K_2 d_2^2 + \frac{1}{2}\, K_3 d_3^2 + \frac{1}{2}\, K_4 d_4^2 + \frac{1}{2}\, K_5 d_5^2 \end{aligned} \tag{2}$$

where $d_1 = Z_2 + \frac{L}{12} Z_3 - Z_1 \qquad d_2 = Z_4 - Z_2 - \frac{L}{3} Z_3$

$d_3 = Z_5 - Z_2 + \frac{2L}{3} Z_3 \qquad d_4 = Z_4 - f_2(t)$

$d_5 = Z_5 - f_1(t)$

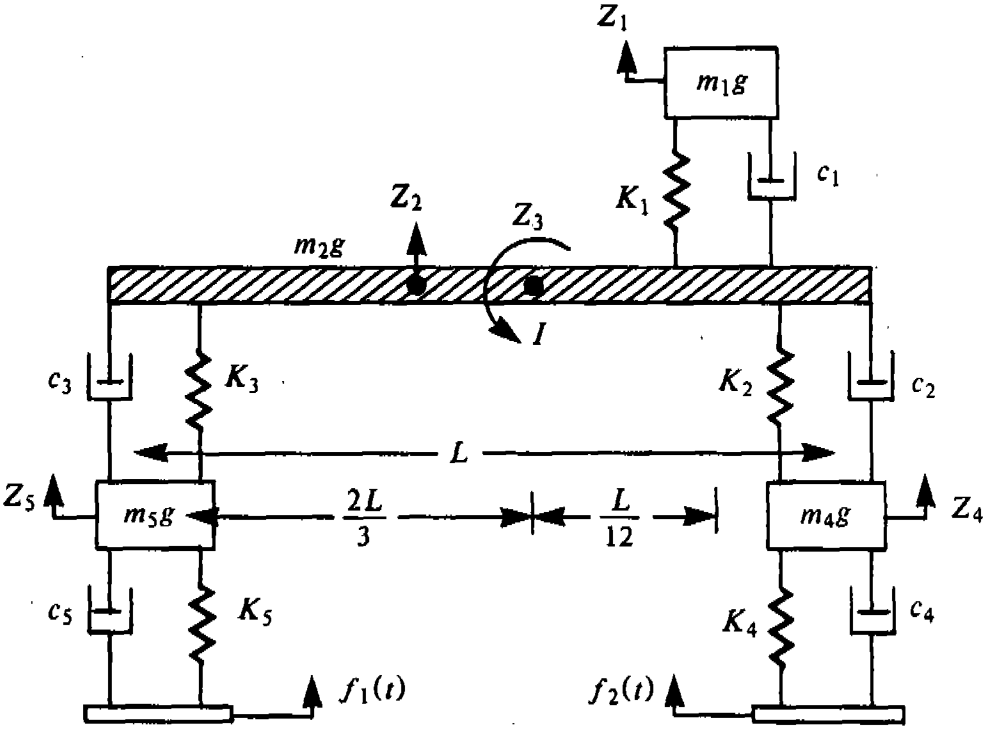


Figure 1.

The Q_i are found from

$$\delta W = \sum_{i=1}^{5} Q_i \delta Z_i = - c_1 \dot{d}_1 \delta d_1 - c_2 \dot{d}_2 \delta d_2 \cdots - c_5 \dot{d}_5 \delta d_5 \qquad (3)$$

The MACSYMA program and output for the above problem is given in its entirety in Appendix I. We comment on the key commands. (C2) establishes the vector Q of generalized coordinates Z1, · · · Z5. (C3) establishes the dependence of the elements of Q on time. (C4)-(C7) establish mass, spring, damping and displacement vectors [for DISP equations following (2) above]. (C11),(C12) derive the generalized forces Qn (=QQ(n) in the program) by picking out the coefficients of δZ_n (=DEL(Q(N)) in the program) in δW in (3) above (=DW in the program). (C9),(C10) establish kinetic and potential energies defined in (2) above. (C13) forms and displays the equations of motion by evaluating (1) above. The key command here is DIFF(EXPR,T) where EXPR is some function of T, which takes the derivative of EXPR with respect to T. Thus

$$\text{DIFF(DIFF(TT,DIFF(Q(N),T)),T)} \equiv \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_n}\right)$$

As requested, the computer then displays the equation, of which we have shown only the first, namely

$$m_1 \ddot{Z}_1 + c_1 \dot{Z}_1 - c_1 \dot{Z}_2 - \frac{L}{12} c_1 \dot{Z}_3 + K_1 Z_1 - K_2 Z_2 - \frac{L}{12} K_1 Z_3 = 0$$

2. SLIDER CRANK PROBLEM

This example illustrates the derivation of equations of motion when constraints are present. The appropriate Lagrange equations are:

$$\frac{\partial}{\partial t}\left[\frac{\partial T}{\partial \dot{q}_i}\right] - \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i} + \left[\frac{\partial \Phi}{\partial q_i}\right]^T \lambda = Q_i \qquad i=1, \cdots, n \tag{4}$$

where T, V, and Q_i are as previously defined and the constraints are represented by k algebraic equations:

$$\Phi(q) = 0$$

and λ are k values of Lagrange multipliers or undetermined coefficients.

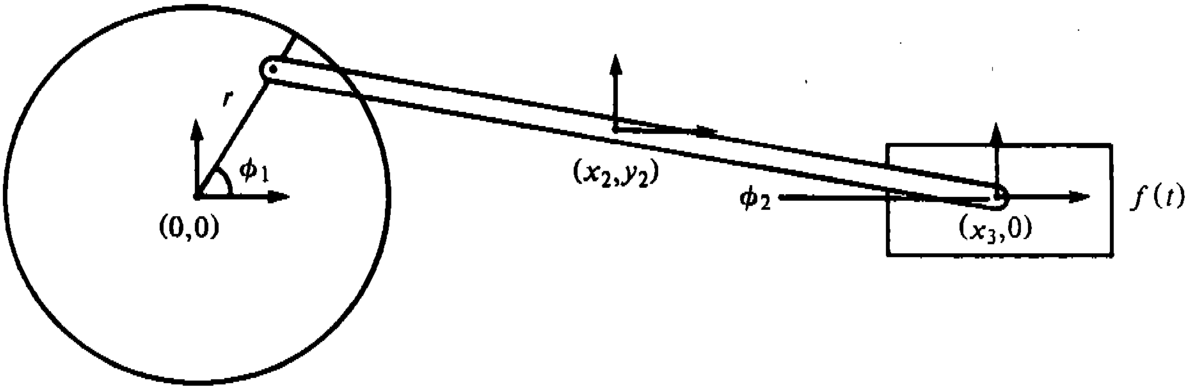


Figure 2.

For the slider crank mechanism shown in Figure 2 we have

$$q = [\phi_1, x_2, y_2, \phi_2, x_3]^T$$

$$T = \frac{1}{2} J_1\dot{\phi}_1^2 + \frac{1}{2} m_2(\dot{x}_2^2 + \dot{y}_2^2) + \frac{1}{2} J_2\dot{\phi}_2^2 + \frac{1}{2} m_3\dot{x}_3^2$$

$$\delta W = f(t)\,\delta x_3$$

$$\begin{aligned}
\Phi_1 &\equiv r\sin\phi_1 - y_2 - l\sin\phi_2 = 0 \\
\Phi_2 &\equiv r\cos\phi_1 - (x_2 - l\cos\phi_2) = 0 \\
\Phi_3 &\equiv x_2 + l\cos\phi_2 - x_3 = 0 \\
\Phi_4 &\equiv y_2 + l\sin\phi_2 = 0
\end{aligned} \tag{5}$$

Again the procedure outlined is easily handled by MACSYMA (see the symbolic program given in Appendix II). From the output of this program we have the following equations of motion for the system:

$$\begin{aligned}
& J_1\ddot{\phi}_1 + [\lambda_1 r\cos\phi_1 - \lambda_2 r\sin\phi_1] = 0 \\
& m_2\ddot{x}_2 + [-\lambda_2 + \lambda_3] = 0 \\
& m_2\ddot{y}_2 + [-\lambda_1 + \lambda_4] = 0 \\
& J_2\ddot{\phi}_2 + [\lambda_1 l\cos\phi_2 - \lambda_2 l\sin\phi_2 - \lambda_3 l\sin\phi_2 + \lambda_4 l\cos\phi_2] = 0 \\
& m_3\ddot{x}_3 - \lambda_3 = f(t)
\end{aligned} \tag{6}$$

Note that (5),(6) are a differential-algebraic system consisting of nine equations in nine unknowns.

This problem has only a single degree of freedom. If this is chosen as ϕ_1, we find from (5) that x_2, y_2, ϕ_2, and x_3 can be expressed in terms of ϕ_1. If these are substituted in the kinetic energy (=TT) expression we find $TT = TT(\dot{\phi}_1,\phi_1)$. The equation of motion is now given by

$$\frac{d}{dt}\left(\frac{\partial TT}{\partial \dot{\phi}_1}\right) - \frac{\partial TT}{\partial \phi_1} = Q_1 \tag{7}$$

i.e., one single differential equation in one unknown, with no constraints. A MACSYMA program for deriving this equation is given in Appendix III.

Equation (7) must be equivalent to nine equations in (5),(6), though derived independently. To deduce (7) directly from (6) we can proceed as follows:

Write (6) and (5) in matrix notation as

$$M\ddot{q} + A^T(q)\lambda = f \tag{8}$$

$$\Phi(q) = 0 \tag{9}$$

$A(q)$ is a 4×5 matrix, so that the equation $A(q)x = 0$ has a solution of the form $x = Cx_0(q)$, where C is an arbitrary constant. Multiplying (8) by x_0^T obtaining

$$x_0^T M\ddot{q} = x_0^T f$$

expresses $x_0(q)$ in terms of ϕ_1 only. This leads to the single differential equation given by (7).

In this problem, another approach would be to choose two generalized coordinates with one side constraint. This would lead to two ordinary differential equations involving ϕ_1, ϕ_2 and one Lagrange's multiplier and one side constraint. These can be obtained either directly [as (7)] or by eliminating three of the λ's in (6).

3. JACOBIANS

If, instead of looking at specific examples as in the last two sections, we consider general formulations, then the following type of situation arises. Suppose that cartesian x-components x_i $(i=1, \cdots, n)$ depend on generalized coordinates q_j $(j=1, \cdots, p)$. Then

$$\dot{x}_i = \sum_{j=1}^{p} \frac{\partial x_i}{\partial q_j}\dot{q}_j$$

$$\ddot{x}_i = \sum_{j=1}^{p} \frac{\partial x_i}{\partial q_j}\ddot{q}_j + \sum_{j=1}^{p}\sum_{k=1}^{p} \frac{\partial^2 x_i}{\partial q_j \partial q_k}\dot{q}_j\dot{q}_k$$

The quantity $d\Phi/dq$ in (4) has the same form as the Jacobian $[\partial x_i/\partial q_j]$ occurring above, and the MACSYMA command for $\partial\phi/\partial q$ is given in the last line of Appendix II.

It is convenient to use the MACSYMA subroutine, or BLOCK to obtain $\partial^2 x_i/\partial q_j \partial q_k$, and this is done in Appendix IV.

4. SENSITIVITY ANALYSIS

The objective of this section is to illustrate how MACSYMA deals with some aspects of sensitivity analysis, with particular reference to the paper by Haug and Ehle [7]. Consider a dynamical system described by design variables $b = [b_1, \cdots, b_k]^T$ and a state variable $z(t) = [z_1(t), \cdots, z_k(t)]^T$ which is the solution of an initial value problem of the form

$$\dot{z} = f(z,b), \qquad 0 < t < T$$
$$z(0) = h(b)$$

where $\dot{z} = dz/dt$ and T is determined by the condition

$$\Lambda(T, z(T)) = 0$$

A typical function that may arise in a design formulation is

$$\psi = g(z(T),b) + \int_0^T F(t,z,b)\,dt \tag{10}$$

It is required to find $d\psi/db$, which is a k-vector. This is done by considering an adjoint variable λ satisfying

$$\dot{\lambda} + f_z^T \lambda = F_z^T$$

and then

$$\frac{d\psi}{db} = g_b - \lambda^T(0) h_b + \int_0^T (F_b - \lambda^T f_b)\,dt \tag{11}$$

The procedure outlined above is carried out by the MACSYMA procedure given in Appendix V. This also illustrates the MACSYMA solution of linear equations by Laplace transform. Consider an example given in Ref. [7], namely a simple oscillator governed by the equation

$$\begin{aligned} &\ddot{x} + kx = 0 \qquad 0 < t < \pi/2 \\ &x(0) = 0 \qquad \dot{x}(0) = v \end{aligned} \tag{12}$$

with $\psi = x(\pi/2)$, $b = [k, v]^T = [b_1, b_2]^T$.

The results of the MACSYMA procedure give the first derivative of the functional ψ as

$$\frac{d\psi}{db} = \begin{bmatrix} -\dfrac{b_2}{2b_1^{3/2}} \sin \dfrac{\pi\sqrt{b_1}}{2} + \dfrac{\pi b_2}{4b_1} \cos \dfrac{\pi\sqrt{b_1}}{2} \\ \dfrac{1}{\sqrt{b_1}} \sin \dfrac{\pi\sqrt{b_1}}{2} \end{bmatrix}^T \tag{13}$$

Higher-order sensitivity analysis requires the Jacobian for which a BLOCK MACSYMA command is given in Appendix IV as discussed in the last section.

Note: MACSYMA is awkward for differentiating functions having a definite integral; e.g., from (10) we have

$$\frac{d\psi}{db} = g_z[z_b(T) + \dot{z}(T)(T)_b] + g_b + \int_0^T [F_z z_b + F_b]\,dt + F(T)(T)_b \tag{14}$$

However, MACSYMA does not perform the derivative under the integral sign (a possible dialogue with MACSYMA is given in Appendix VI).

5. A SPACECRAFT PROBLEM

Levinson [10] has described in detail an application of the symbolic language FORMAC to formulate the spacecraft problem shown in Figure 3, consisting of two rigid bodies with a common axis of rotation b.

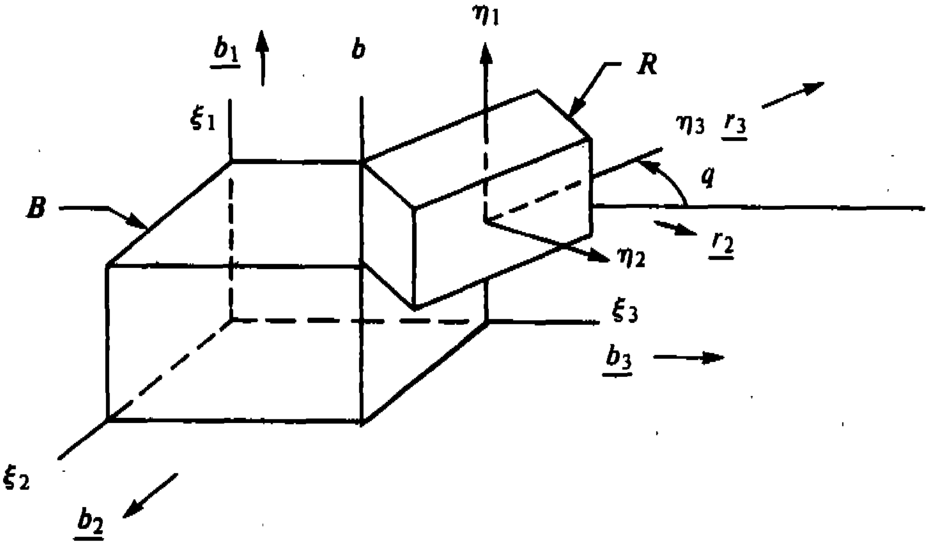


Figure 3.

The equations are given in Ref. [10] in complete detail, and are translated into MACSYMA in Appendix VII. To illustrate the point we give typical equations with MACSYMA equivalents:

Equations from Ref. [10]			MACSYMA
$\underline{r}_2$	$= \cos q\, \underline{b}_2 + \sin q\, \underline{b}_3$	(1)	`R[2]:COS(Q)*B[2]+SIN(Q)*B[3];`
ω^B	$= u_1 \underline{b}_1 + u_2 \underline{b}_2 + u_3 \underline{b}_3$	(3)	`WB:U[1]*B[1]+U[2]*B[2]+U[3]*B[3];`
u_4	$= \dot{q}$		`U[4]:DIFF(Q,T);`
α^R	$= \frac{d}{dt}(\omega^R) + \omega^B \times \omega^R$	(7)	

We could implement this last mathematical expression (7) [10] by converting the vector product into matrix form but it was simpler to write a BLOCK function to do this, as in Appendix VII. The MACSYMA equivalent of (7) is now

```
ALPR:DIFF(WR,T)+CROSS(WB,WR);
```

We discuss only one other correspondence. Equation (27) in Ref. [10] is

$$F_r = \frac{\partial v^{B^*}}{\partial u_r} \cdot (F)_B + \frac{\partial \omega^B}{\partial u_r} \cdot (T)_B \qquad (r=1, \cdots, 7)$$

becomes in MACSYMA

```
F(R):=DOT(DIFF(VBS,U[R]),FB)
        +DOT(DIFF(WB,U[R]),TB);
```

Dot in the above is defined by another block in Appendix VII. The distinction between := and :, as used in this command, is discussed in Appendix B.

The complete set of equations given in Ref. [10] is generated by Appendix VII. The reader should compare the corresponding FORMAC program given in Levinson [10].

6. AN EXAMPLE OF MANIPULATION AND SIMPLIFICATION USING MACSYMA

In previous sections we have not found it necessary to use powerful commands in MACSYMA concerned with the simplification of complicated expressions. As an introduction to the manipulation needed in the later sections, we present the following simple dialog:

(C2) F:(X+Y+Z)^2/Y;

(D2) $$\frac{(Z+Y+X)^2}{Y}$$

(C3) EXPAND(%);

(D3) $$\frac{Z^2}{Y} + \frac{2XZ}{Y} + 2Z + Y + \frac{X^2}{Y} + 2X$$

(C4) COMBINE(%);

(D4) $$\frac{Z^2+2XZ+X^2}{Y} + 2Z + Y + 2X$$

(C5) XTHRU(%);

(D5) $$\frac{Z^2+Y(2Z+Y+2X)+2XZ+X^2}{Y}$$

(C6) RATSIMP(%);

(D6) $$\frac{Z^2+(2Y+2X)\,Z+Y^2+2XY+X^2}{Y}$$

(C7) EV(%,Z=0);

(D7) $$\frac{Y^2+2XY+X^2}{Y}$$

(C8) SUBST(SIN(2*TH),X,%);

(D8) $$\frac{Y^2+2\,SIN(2TH)\,Y+SIN^2(2TH)}{Y}$$

(C9) TRIGEXPAND(%);

(D9) $$\frac{Y^2+4\,COS(TH)\,SIN(TH)\,Y+4\,COS^2(TH)\,SIN^2(TH)}{Y}$$

(C10) TRIGREDUCE(%);

(D10) $$Y - \frac{COS(4TH)}{2Y} + \frac{1}{2Y} + 2\,SIN(2TH)$$

7. THE FOUR-BAR LINKAGE COUPLER CURVE

The objective of this section is to illustrate how MACSYMA performs algebraic and trigonometric manipulations encountered in the analysis and synthesis of linkages.

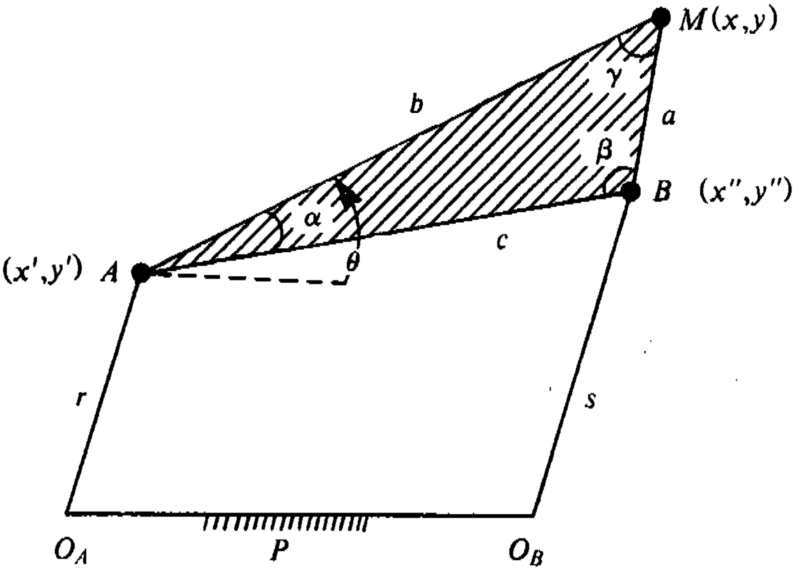


Figure 4.

Consider first, following Hartenberg and Denavit [5] (p. 150), the four-bar linkage AO_AO_BB shown in Figure 4, where the bars lie in a plane and are pin-jointed at A, O_A, O_B, and B, and the positions of O_A and O_B are fixed. MAB is a lamina, so that M is fixed relative to A and B. If BO_B is rotated about O_B, the point M will trace a planar curve, the equation of which we wish to determine.

Using the linkage parameter shown in Figure 4 we have:

$$\begin{aligned} x' &= x - b\cos\theta & x'' &= x - a\cos(\theta+\gamma) \\ y' &= y - b\sin\theta & y'' &= y - z\sin(\theta+\gamma) \\ r^2 - x'^2 - y'^2 &= 0 & s^2 - (x''-p)^2 - y''^2 &= 0 \end{aligned} \tag{15}$$

The required equation for the motion of $M(x,y)$ is obtained by eliminating (x',y'), (x'',y''), and θ from (15). This is done in Appendix VIII. Elimination of (x',y') and (x'',y'') leads to the equation of the form

$$\begin{aligned} N\cos\theta - L\sin\theta &= \phi \\ -P\cos\theta + M\sin\theta &= \psi \end{aligned} \tag{16}$$

where $L = (x-p)\sin\gamma - y\cos\gamma$, $\quad N = (x-p)\cos\gamma + y\sin\gamma$

$$M = y\,, \quad P = -x$$

$$\phi = \frac{1}{B}\,(y^2+x^2-r^2-b^2)\,, \quad \psi = \frac{1}{2A}\left[(x-p)^2+y^2+a^2-s^2\right]$$

Eliminating θ from (16) gives

$$(P\psi+N\phi)^2 + (M\psi+L\phi)^2 = (LP-MN)^2 \tag{17}$$

This sixth-degree polynomial in (x,y) is called the tricircular sextic. The determinant of (16) vanishes when $LP-NM = 0$, i.e.,

$$x(x-p) + y^2 - py\cot\gamma = 0 \tag{18}$$

The above equation is called a circle of singular foci.

We next derive a basic relation used to synthesize four-bar linkages, namely the so-called displacement equation which gives the output angle ψ for a given input angle ϕ in Figure 5.

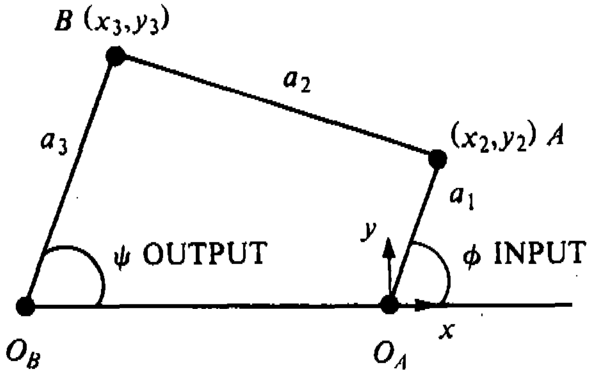


Figure 5.

$$\begin{aligned} &x_2 = a_1\cos\phi\,, \qquad y_2 = a_1\sin\phi \\ &x_3 = -a_4 + a_3\cos\psi\,, \qquad y_3 = a_3\sin\psi \\ &a_2^2 - (x_2 - x_3)^2 - (y_2 - y_3)^2 = 0 \end{aligned} \tag{19}$$

Eliminating (x_2,y_2) and (x_3,y_3) from (19) leads to

$$A\sin\psi + B\cos\psi = C \tag{20}$$

where $A = 2a_1a_3\sin\phi\,, \qquad B = 2a_1a_3\cos\phi + 2a_3a_4$

$$C = 2a_1a_4\cos\phi + (a_4^2 + a_3^2 - a_2^2 + a_1^2)$$

To solve (20) for ψ set

$$\sin\psi = \frac{2\tan\psi/2}{1+\tan^2\psi/2} \qquad \cos\psi = \frac{1-\tan^2\psi/2}{1+\tan^2\psi/2} \tag{21}$$

Substitution into (20) leads to a quadratic in $\tan\psi$, the solution of which is

$$\tan\frac{\psi}{2} = \frac{1}{(B+C)}\left\{A \pm \sqrt{A^2+B^2-C^2}\right\}$$

The two solutions correspond to the two ways to close the four-bar linkage shown in Figure 6.

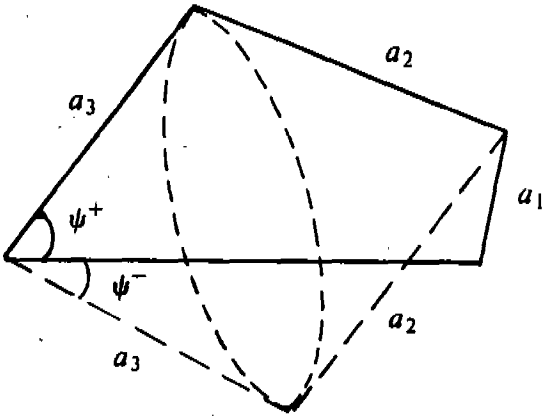


Figure 6.

For the purpose of synthesis, (20) can be rewritten as:

$$K_1\cos\phi - K_2\cos\psi + K_3 = \cos(\phi - \psi) \tag{22}$$

$$\text{with} \quad K_1 = \frac{a_4}{a_3}, \quad K_2 = \frac{a_4}{a_1}, \quad K_3 = a_2^2 = (a_1^2 + a_3^2 + a_4^2 - 2a_1a_3K_3)/a_1a_3 \tag{23}$$

Hartenberg and Denavit [5] (p. 297) discuss the problem of designing a planar four-bar linkage such that, to three given positions ϕ_1, ϕ_2, and ϕ_3 of the crank O_AA, there correspond three prescribed positions ψ_1, ψ_2, and ψ_3 of the follower O_BB. The form of (22) is well-suited for this purpose. The solution in this case is obtained by solving the set of three simultaneous equations for K_1, K_2, and K_3 obtained by substituting $\phi = \phi_i$, $\psi = \psi_i$, $i = 1, 2, 3$ in (22), and then obtaining a_3, a_1, and a_2 from (23) (a_4 can be selected equal to one). Appendices VIII and IX give the MACSYMA program to carry out the above procedures.

8. DUAL-NUMBER QUATERNIONS

Next, consider a laborious calculation contained in the appendix to the Yang and Freudenstein paper [19] in connection with the analysis of a spatial four-bar mechanism. We are given

$$A(\hat{\theta}_1)\sin\hat{\theta}_4 + B(\hat{\theta}_1)\cos\hat{\theta}_4 = C(\hat{\theta}_1) \tag{24}$$

where
$$\begin{aligned} A(\hat{\theta}_1) &= \sin\hat{\alpha}_{12}\sin\hat{\alpha}_{34}\sin\hat{\theta}_1 \\ B(\hat{\theta}_1) &= -\sin\hat{\alpha}_{34}(\sin\hat{\alpha}_{41}\cos\hat{\alpha}_{12} + \cos\hat{\alpha}_{41}\sin\hat{\alpha}_{12}\cos\hat{\theta}_1) \\ C(\hat{\theta}_1) &= \cos\hat{\alpha}_{23} - \cos\hat{\alpha}_{34}(\cos\hat{\alpha}_{41}\cos\hat{\alpha}_{12} - \sin\hat{\alpha}_{41}\sin\hat{\alpha}_{12}\cos\hat{\theta}_1) \end{aligned} \tag{25}$$

Here
$$\begin{aligned} \hat{\alpha}_{12} &= \alpha_{12} + \epsilon a_{12}\,, & \hat{\theta}_1 &= \theta_1 + \epsilon s_{11} \\ \hat{\alpha}_{23} &= \alpha_{23} + \epsilon a_{23}\,, & \hat{\theta}_2 &= \theta_2 + \epsilon s_2 \\ \hat{\alpha}_{34} &= \alpha_{34} + \epsilon a_{34}\,, & \hat{\theta}_3 &= \theta_3 + \epsilon s_3 \\ \hat{\alpha}_{41} &= \alpha_{41} + \epsilon a_{41}\,, & \hat{\theta}_4 &= \theta_4 + \epsilon s_4 \end{aligned} \tag{26}$$

where ϵ is a symbol with the property $\epsilon^2 = 0$. This implies that, if $\hat{\theta} = \theta + \epsilon s$, then

$$\sin\hat{\theta} = \sin\theta + \epsilon s\cos\theta\,, \qquad \cos\hat{\theta} = \cos\theta - \epsilon s\sin\theta\,.$$

It is then clear that (24) can be reduced to the form

$$P + \epsilon Q = R + \epsilon S \tag{27}$$

where P, Q, R, and S are independent of ϵ. It is required to find the explicit form of P, Q, R, and S. To calculate this by hand is extremely laborious, but straightforward in MACSYMA. The program is given in Appendix X.

Three-dimensional problems in kinematics and dynamics involve laborious calculations involving Euler angles and Euler parameters. (See, for instance, Nikravesh, et al. [13], and Wittenburg [18].) These calculations are easily handled in MACSYMA in a routine fashion. The techniques involved are illustrated in connection with other examples in this paper, so we do not elaborate further.

9. ROBOT ARMS – THE DIRECT PROBLEM

Robot arm manipulators can be considered to consist of a series of links connected together by joints. It is convenient to use cartesian coordinates by assigning a separate coordinate frame to each link. Without going into detail (which can be found in Paul [15], for instance), the relation between the coordinate frames assigned to one link and the next, consisting of translations and rotations, can be derived by a 4×4 matrix of the form

$$A = \begin{bmatrix} n_x & o_x & a_x & p_x \\ n_y & o_y & a_y & p_y \\ n_z & o_z & a_z & p_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{28}$$

where the elements of the top left 3×3 submatrix are the direction cosines representing the rotations and (p_x, p_y, p_z) is the translation.

The position and orientation of the coordinate frame of the end of the manipulator is specified by six parameters (3 translations, 3 rotations). A general manipulator can be designed using six links, each having one degree of freedom. If T_6 denotes the A-matrix corresponding to the end of the manipulator, and A_i $(i=1, \cdots, 6)$ are the A-matrices for the individual links, T_6 is given terms of the A_i by

$$T_6 = A_1 A_2 A_3 A_4 A_5 A_6$$

A typical A-matrix for a link is

$$A_2 = \begin{bmatrix} \cos\theta_2 & 0 & \sin\theta_2 & 0 \\ \sin\theta_2 & 0 & -\cos\theta_2 & 0 \\ 0 & 1 & 0 & d_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

If d_2 is fixed and θ_2 is a variable representing rotation of the second link, this is called a *revolute* joint. If θ_2 is fixed and the translation d_2 is varying, this is called a *prismatic* joint.

The so-called "direct" problem is: given the A_i, find T_6—this is obviously straightforward, although algebraically laborious (see Paul [15], p. 59 and Appendix X).

The main computational problem connected with the direct problem is the question of differential motion discussed in Paul [15] (Chapter 4). These are important in connection with dynamic analysis of manipulators, sensitivity analyses, and small adjustments of the end manipulator.

Without going into detail (which can be found in Ref. [15]), the computational problem involved is the following. Suppose that the six parameters representing degrees of freedom are denoted by a 6-vector x, small changes in these parameters are denoted by Δx, and the corresponding small changes in the three displacements and three rotational parameters of the end point frame of the manipulator are denoted by the 6-vector A, then we have a relation of the form

$$\Delta A = J \Delta x$$

where the i^{th} column of J is $\begin{bmatrix} d_i \\ \delta_i \end{bmatrix}$; where, for $i=1, \cdots, 6$, and for a *revolute* joint,

$$d_i = \begin{bmatrix} (-n_{ix}p_{iy} + n_{iy}p_{ix}) \\ (-o_{ix}p_{iy} + o_{iy}p_{ix}) \\ (-a_{ix}p_{iy} + a_{iy}p_{ix}) \end{bmatrix} \qquad \delta_i = \begin{bmatrix} n_z \\ o_z \\ a_z \end{bmatrix}$$

and for a *prismatic* joint

$$d_i = \begin{bmatrix} n_z \\ o_z \\ a_z \end{bmatrix} \qquad \delta_i = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The MACSYMA program for the symbolic computation of A and the numerical example in Ref. [15] are given in Appendix XI (see Ref. [15], pp. 104-107).

10. ROBOT ARMS – THE INVERSE PROBLEM

The "inverse" problem consists of obtaining the A_i, $i=1, \cdots, 6$, given numerical values of T_6. In theory this can be done from

$$A_1 A_2 A_3 A_4 A_5 A_6 = T_6 \tag{29}$$

which gives 12 equations in 6 unknowns (the 6 degrees of freedom of the links). These equations are redundant, and there are only six independent equations in the six unknowns. However, the equations are highly complicated. The method used in practice is to consider also the following equations which are completely equivalent to (29):

$$\left.\begin{aligned} A_2 A_3 A_4 A_5 A_6 &= A_1^{-1} T_6 \\ A_3 A_4 A_5 A_6 &= A_2^{-1} A_1^{-1} T_6 \\ A_4 A_5 A_6 &= A_3^{-1} A_2^{-1} A_1^{-1} T_6 \\ A_5 A_6 &= A_4^{-1} A_3^{-1} A_2^{-1} A_1^{-1} T_6 \\ A_6 &= A_5^{-1} A_4^{-1} A_3^{-1} A_2^{-1} A_1^{-1} T_6 \end{aligned}\right\} \tag{30}$$

[In practice the A_i^{-1} are usually easily obtained from the A_i.] Equations (29) and (30) give 72 equations for the 6 unknowns. The procedure is now to pick out the simplest 6 independent equations from the set of 72. The simplest solution occurs when one of the equations involves only one unknown, say x_1, another equation involves x_1 and a second unknown x_2, a third equation involves only x_1, x_2, x_3, and so on. The system can then be solved sequentially. This is the solution with Stanford and the elbow manipulators described by Paul [15].

A more complicated situation occurs in the robot arm discussed by Lumelsky [11], where such a simple sequence of equations cannot be found. Instead, the simplest set is of the form

$$\begin{aligned} x_1 &= f_1(x_3, x_4) \qquad x_2 = f_2(x_3, x_4) \\ x_3 &= f_3(x_1, x_2) \qquad x_4 = f_4(x_1, x_2) \end{aligned} \tag{31}$$

These can be solved by straightforward iteration.

In Appendix XII we give a MACSYMA program for selecting the basic 6 equations from the 72 available. This can be done automatically by using the command FREEOF to print out a dependency table showing which of the variables occur in each of the 72 equations. Equation (31) can be deduced directly from this dependency table.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the help of Maria Barnum in preparing this difficult manuscript for publication.

MACSYMA work of the Mathlab group is currently supported, in part, by the National Aeronautics and Space Administration under grant NSG 1323, by the Office of Naval Research under grant N00014-77-C-0641, by the U.S. Department of Energy under grant ET-78-C-02-4687, and by the U.S. Air Force under grant F49620-79-C-020. Funding for one of the authors [B.N.] was provided by the Army Research Office.

REFERENCES

[1] L. Bordoni and A. Golarossi, "An Application of Reduce to Industrial Machinery," *ACM Sigsam Bull. No. 58,* 115, 8-12 (1981).

[2] O. Bottema and B. Roth, *Theoretical Kinematics,* North-Holland, 1979.

[3] R.D. Drinkard and N.K. Sulinski, *MACSYMA: A Program for Computer Algebraic Manipulations,* Naval Underwater Systems Center, Newport, Rhode Island, NUSC Tech. Doc. 6401, 10 March 1981.

[4] H. Goldstein, *Classical Mechanics,* Addison-Wesley, 1959.

[5] R.S. Hartenberg and J. Denavit, *Kinematic Synthesis of Linkages,* McGraw-Hill, 1964.

[6] E.J. Haug and J.S. Arora, *Applied Optimal Design,* Wiley, 1979.

[7] E.J. Haug and P.E. Ehle, "Second-Order Design Sensitivity Analysis of Mechanical System Dynamics," *Int. J. Num. Meth. Eng. 18,* 1699-1717 (1982).

[8] M.A. Hussain and B. Noble, "Application of MACSYMA to Calculations in Applied Mathematics," General Electric Company Report No. 83CRD054, March 1983.

[9] E.J. Kreuzer, "Dynamical Analysis of Mechanisms Using Symbolical Equation Manipulation," *Proc. Fifth World Congress on Theory of Machines and Mechanisms,* ASME (1979).

[10] D.A. Levinson, "Equations of Motion for Multiple-Rigid-Body Systems via Symbolic Manipulation," *J. Spacecraft and Rockets 14,* 479-487 (1977).

[11] V.J. Lumelsky, "Iterative Procedure for Inverse Coordinate Transformation for One Class of Robots," General Electric Company Report No. 82CRD332, February 1983.

[12] *MACSYMA: The Reference Manual,* Version 10, 1983, Math Lab Group, Laboratory for Computer Science, MIT. *See also MACSYMA Primer.*

Information, plus the MACSYMA tape (available to colleges and universities at special rates), is available from: Symbolics Inc., 257 Vassar St., Cambridge, MA 02139.

[13] P.E. Nikravesh, R.A. Wehage, and E.J. Haug, *Computer Aided Analysis of Mechanical Systems* (1982), to be published.

[14] B. Paul, *Kinematics and Dynamics of Planar Machinery,* Prentice-Hall, 1979.

[15] R.P. Paul, *Robot Manipulators,* MIT Press, 1981.

[16] W.O. Schiehlen and E.J. Kreuzer, "Symbolic Computerized Derivation of Equations of Motion," in *Dynamics of Multibody Systems 1,* IUTAM Conf. (K. Magnus, ed.), Springer-Verlag, 1978, pp. 290-305.

[17] M. Vukobratovic and V. Potkonjak, *Dynamics of Manipulation Robots,* Springer-Verlag, 1982.

[18] J. Wittenberg, *Dynamics of Systems of Rigid Bodies,* Teubner, 1977.

[19] A.T. Yang and F. Freudenstein, "Application of Dual-Number Quaternion Algebra to the Analysis of Spatial Mechanisms," *Trans. ASME, J. Appl. Mech.,* 300-308 (1964).

APPENDIX I

```
(C2)  Q:[Z1,Z2,Z3,Z4,Z5];
```

(D2) $$[Z1, Z2, Z3, Z4, Z5]$$

```
(C3)  DEPENDS(Q,T);
```

(D3) $$[Z1(T), Z2(T), Z3(T), Z4(T), Z5(T)]$$

```
(C4)  MASS:[M1,M2,I,M4,M5];
```

(D4) $$[M1, M2, I, M4, M5]$$

```
(C5)  SPRING:[K1,K2,K3,K4,K5];
```

(D5) $$[K1, K2, K3, K4, K5]$$

```
(C6)  DASH:[CC1,CC2,CC3,CC4,CC5];
```

(D6) $$[CC1, CC2, CC3, CC4, CC5]$$

```
(C7)  DISP:[Z2+L/12*Z3-Z1,Z4-Z2-L/3*Z3,Z5-Z2+2*L/3*Z3,Z4-F1(T),Z5-F2(T)];
```

(D7) $$\left[\frac{LZ3}{12}+Z2-Z1,\ Z4-\frac{LZ3}{3}-Z2,\ Z5+\frac{2LZ3}{3}-Z2,\ Z4-F1(T),\ Z5-F2(T)\right]$$

```
(C8)  DERIVABBREV:TRUE;
```

(D8) $$TRUE$$

```
(C9)  TT:1/2*(MASS . DIFF(Q,T)^2);
```

(D9) $$\frac{M5(Z5_T)^2+M4(Z4_T)^2+I(Z3_T)^2+M2(Z2_T)^2+M1(Z1_T)^2}{2}$$

```
(C10) POT:1/2*(SPRING . DISP^2);
```

(D10) $$\left(K3\left(Z5+\frac{2LZ3}{3}-Z2\right)^2+K5\left(Z5-F2(T)\right)^2+K2\left(Z4-\frac{LZ3}{3}-Z2\right)^2+K4\left(Z4-F1(T)\right)^2+K1\left(\frac{LZ3}{12}+Z2-Z1\right)^2\right)/2$$

```
(C11) DW:-DIFF(DISP,T)*DIFF(DISP);
```

(D11) $$\begin{aligned}\Bigg[&-\left(\frac{LZ3_T}{12}+Z2_T-Z1_T\right)\left(DEL(Z2)+\frac{L\,DEL(Z3)}{12}-DEL(Z1)\right.\\ &\left.+\left(\frac{LZ3_T}{12}+Z2_T-Z1_T\right)DEL(T)+\frac{Z3\,DEL(L)}{12}\right),\\ &-\left(Z4_T-\frac{LZ3_T}{3}-Z2_T\right)\left(DEL(Z4)-DEL(Z2)-\frac{L\,DEL(Z3)}{3}\right.\\ &\left.+\left(Z4_T-\frac{LZ3_T}{3}-Z2_T\right)DEL(T)-\frac{Z3\,DEL(L)}{3}\right),\\ &-\left(Z5_T+\frac{2LZ3_T}{3}-Z2_T\right)\left(DEL(Z5)-DEL(Z2)+\frac{2L\,DEL(Z3)}{3}\right.\\ &\left.+\left(Z5_T+\frac{2LZ3_T}{3}-Z2_T\right)DEL(T)+\frac{2Z3\,DEL(L)}{3}\right),\\ &-\left(Z4_T-F1(T)_T\right)\left(DEL(Z4)+\left(Z4_T-F1(T)_T\right)DEL(T)\right),\\ &-\left(Z5_T-F2(T)_T\right)\left(DEL(Z5)+\left(Z5_T-F2(T)_T\right)DEL(T)\right)\Bigg]\end{aligned}$$

```
(C12) FOR N THRU 5 DO QQ[N]:RATCOEF(DW,DEL(Q[N]));
```

(D12) $$DONE$$

```
(C13) FOR N THRU 5 DO LDISPLAY(EQUATION[N]
      = DIFF(DIFF(TT,DIFF(Q[N],T)),T)-DIFF(TT,Q[N])+DIFF(POT,Q[N])-DASH . QQ[N]);
```

(E13) $$EQUATION_1=-K1\left(\frac{LZ3}{12}+Z2-Z1\right)+M1Z1_{TT}-\frac{CC1\,(LZ3_T+12Z2_T-12Z1_T)}{12}$$

(E14) $$\begin{aligned}EQUATION_2=&\frac{-K3\left(Z5+\frac{2LZ3}{3}-Z2\right)-2K2\left(Z4-\frac{LZ3}{3}+2K1\left(\frac{LZ3}{12}+Z2-Z1\right)\right)}{2}\\ &+M2Z2_{TT}-\frac{CC3\,(3Z5_T+2LZ3_T-3Z2_T)}{3}-\frac{CC2\,(3Z4_T-LZ3_T-3Z2_T)}{3}\\ &+\frac{CC1\,(LZ3_T+12Z2_T-12Z1_T)}{12}\end{aligned}$$

(E15) $$\begin{aligned}EQUATION_3=&\frac{\dfrac{4K3L\left(Z5+\frac{2LZ3}{3}-Z2\right)}{3}-\dfrac{2K2L\left(Z4-\frac{LZ3}{3}-Z2\right)}{3}+\dfrac{K1L\left(\frac{LZ3}{12}+Z2-Z1\right)}{6}}{2}\\ &+IZ3_{TT}-\frac{CC2\,(3LZ4_T-L^2Z3_T-3LZ2_T)}{9}\\ &+\frac{CC3\,(6LZ5_T+4L^2Z3_T-6LZ2_T)}{9}+\frac{CC1\,(L^2Z3_T+12LZ2_T-12LZ1_T)}{144}\end{aligned}$$

(E16) $$\begin{aligned}EQUATION_4=&M4Z4_{TT}+\frac{2K2\left(Z4-\frac{LZ3}{3}-Z2\right)+2K4\left(Z4-F1(T)\right)}{2}\\ &\frac{CC2\,(3Z4_T-LZ3_T-3Z2_T)}{3}-CC4\left(-Z4_T+F1(T)_T\right)\end{aligned}$$

(E17) $$\begin{aligned}EQUATION_5=&M5Z5_{TT}+\frac{2K3\left(Z5+\frac{2LZ3}{3}-Z2\right)+2K5\left(Z5-F2(T)\right)}{2}\\ &\frac{CC3\,(3Z5_T+2LZ3_T-3Z2_T)}{3}-CC5\left(-Z5_T+F2(T)_T\right)\end{aligned}$$

```
/* APPENDIX I (CONT.)   MACSYMA PROGRAM  */

/*.....GENERALIZED COORDINATES...*/
Q:[Z1,Z2,Z3,Z4,Z5];
DEPENDS(Q,T);
/*.....GENERALIZED MASS..........*/
MASS:[M1,M2,I,M4,M5];
/*.....SPRING CONSTANT...........*/
SPRING:[K1,K2,K3,K4,K5];
/*.....DAMPING CONSTANT..........*/
DASH:[CC1,CC2,CC3,CC4,CC5];
/*.....GENERILIZED DISPLACEMENT..*/
DISP:[Z2+L/12*Z3-Z1,Z4-Z2-L/3*Z3,Z5-Z2+2*L/3*Z3,Z4-F1(T),Z5-F2(T)];
DERIVABBREV:TRUE;
/*.....KINETIC AND POTENTIAL ENERGIES..*/
TT:1/2*MASS.DIFF(Q,T)^2;
POT:1/2*SPRING.DISP^2;
/*........DW..........*/
DW:-DIFF(DISP,T)*DIFF(DISP);
FOR N THRU 5 DO QQ[N]:RATCOEF(DW,DEL(Q[N]));
/*.....EQUATIONS OF MOTION............*/
FOR N:1 THRU 5 DO LDISPLAY (EQUATION[N]=
DIFF(DIFF(TT,DIFF(Q[N],T)),T)
-DIFF(TT,Q[N])
+DIFF(POT,Q[N])
-DASH.QQ[N]);
```

APPENDIX II

```
/* CO-ORDINATES */
  Q:[PH1,X2,Y2,PH2,X3];
  DEPENDS(Q,T);
  MASS:[J1,M2,M2,J2,M3];
   CONSTRAINT:[R*SIN(PH1)-Y2+L*SIN(PH2),
        R*COS(PH1)-X2+L*COS(PH2),
        X2+L*COS(PH2)-X3,
        Y2+L*SIN(PH2)];
/* LAGRANGE MULTIPLIERS       */
  LAM:[LAM1,LAM2,LAM3,LAM4];
/* KINETIC ENERGY             */
  TT:(1/2*DIFF(Q,T)^2).MASS;
/* EQUATION OF MOTION         */
  FOR I:1 THRU 5 DO LDISPLAY (QQ[I]=DIFF(DIFF(TT,DIFF(Q[I],T)),T)
       -DIFF(TT,Q[I]) +
      LAM.(DIFF(CONSTRAINT,Q[I])));
```

APPENDIX III

```
/*  ALTERNATE WAY TO DO SLIDER CRANK PROBLEM  */

  Q:[PH1,X2,Y2,PH2,X3];
  DEPENDS(Q,T);
  Y2:1/2*R*SIN(PH1);
  X2:R*COS(PH1)+L*SQRT(1-R**2/(4*L**2)*SIN(PH1));
  X3:X2+L*COS(PH1);
  PH2:ASIN(R/(2*L)*SIN(PH1));
  MASS:[J1,M2,M2,J2,M3];
  DERIVABBREV:TRUE;
  TT:(1/2*DIFF(Q,T)^2).MASS;
  TT:EV(TT,DIFF);
  EQUATION:DIFF(DIFF(TT,DIFF(PH1,T)),T)-DIFF(TT,PH
```

APPENDIX IV

```
/*..............TEST FOR JACOBIAN......HIGHER DERIVATIVES..
  DIMENSION OF A X AND Y ARE P K AND M RESPECTIVELY
    AND SECOND ORDER DERIVATIES ARE FORMED   */
 JAC2(A,P,X,K,Y,M):=BLOCK(
 FOR L:1 THRU P DO (
 DEPENDENT:DETERMINANT(A[L]),
 FOR I:1 THRU K DO (
 VARIABLE1:DETERMINANT(X[I]),
 FOR J:1 THRU M DO (
 VARIABLE2:DETERMINANT(Y[J]),
 PART[L,I,J]:DIFF(DIFF(DEPENDENT,VARIABLE1),VARIABLE2)))));
/*  FOLLOWING IS A SIMPLE EXAMPLE......   */
 A:MATRIX([Z1*B1**2-2*B1*Z3],[B2**2*Z2],[3*B1**2*Z3-B1**3*Z1]);
 X:MATRIX([B1],[B2],[B3]);
 Y:MATRIX([Z1],[Z2],[Z3]);
 JAC2(A,3,X,3,Y,3);
 FOR L:1 THRU 3 DO ( FOR I:1 THRU 3 DO ( FOR J:1 THRU 3 DO
 (LDISPLAY (PART[L,I,J])))));
/*  IN THE FOLLOWING CASE A HAS DIMENSION N BY M
   AND B HAS P BY 1 AND FIRST DERIVATIVES ARE FORMED.  */
 JAC3(A,N,M,B,P):-BLOCK(FOR L:1 THRU P DO (
 VARIABLE:DETERMINANT(B[L]),
 FOR I:1 THRU N DO (
 FOR J:1 THRU M DO (
 DEPENDENT:(A[I,J]),
 PART[L,I,J]:DIFF(DEPENDENT,VARIABLE)))));
 AA:MATRIX([Z1**2*B3,Z2**2*B1,Z3*B1],[Z2**3*B3,Z1*Z2,Z1*B3],
 [Z3**2+Z2,Z2*B2*Z3,Z1*B3**2]);
 JAC3(AA,3,3,X,3);
 FOR L:1 THRU 3 DO (
 FOR I:1 THRU 3 DO (
 FOR J:1 THRU 3 DO LDISPLAY(
 PART[L,I,J])));
```

APPENDIX V

```
/*..THIS IS THE FIRST PROBLEM OF HAUG, SEE [7]
  B IS THE DESIGN VARIABLE AND Z IS STATE VARIABLE
  F IS THE RIGHT HAND SIDE OF DIFFERENTIAL EQUATION
  AND H IS B.C  */
H:MATRIX([0],[B2]);
B:MATRIX([B1],[B2]);
F:MATRIX([Z2(T)],[-B1*Z1(T)]);
Z:MATRIX([Z1(T)],[Z2(T)]);
/*..SET UP DIFF EQ AND SOLVE BY LAP TRANSFORM..*/
/*............................................*/
EQ1:DIFF(Z,T)-F;
/*  INITIAL VALUES ARE GIVEN HERE  */
ATVALUE(Z1(T),T=0,0);
ATVALUE(Z2(T),T=0,B2);
P1:DETERMINANT(EQ1[1]);
P2:DETERMINANT(EQ1[2]);
EQ2:LAPLACE(P1,T,S);
EQ3:LAPLACE(P2,T,S);
LINSOLVE([EQ2,EQ3],[LAPLACE(Z1(T),T,S),
LAPLACE(Z2(T),T,S)]);
PP1:RHS(FIRST(%));
PP2:RHS(LAST(%TH(2)));
/*'INVERSE LAPLACE TRANSFORM  */
Z1(T):=ILT(PP1,S,T);
Z2(T):=ILT(PP2,S,T);
Z1(T);
Z1(T):-"%;
Z2(T);
Z2(T):-"%;
JAC(A,N,B,K):-BLOCK([PART],
FOR I:1 THRU N DO (
DEPENDENT:DETERMINANT(A[I]),
FOR J:1 THRU K DO (
VARIABLE:DETERMINANT(B[J]),
PART[I,J]:DIFF(DEPENDENT,VARIABLE))),
GENMATRIX(PART,N,K,1,1));
/* NOW WE SOLVE FOR ADJOINT VARIABLES */
DEPENDS([LAM1,LAM2],[T]);
LAM:MATRIX([LAM1(T)],[LAM2(T)]);

EQ1:DIFF(LAM,T)+TRANSPOSE(JAC(F,2,Z,2)).LAM;
/* SINCE LAPLACE TRANSFORM SOLVES WITH INITIAL VALUES
   ONLY WE ASUUME ALPHA AND BETA AS THE INITIAL VALUES
   AND SOLVE FOR ALPHA AND BETA FROM THE FINAL VALUES
   OF THE SOLUTON */
ATVALUE(LAM1(T),T=0,ALPHA);
ATVALUE(LAM2(T),T=0,BETA);
P1:DETERMINANT(EQ1[1]);
P2:DETERMINANT(EQ1[2]);
EQ2:LAPLACE(P1,T,S);
EQ3:LAPLACE(P2,T,S);
LINSOLVE([EQ2,EQ3],[LAPLACE(LAM1(T),T,S),
LAPLACE(LAM2(T),T,S)]);
PP1:RHS(FIRST(%));
PP2:RHS(LAST(%TH(2)));
LAM1(T):-ILT(PP1,S,T);
LAM2(T):-ILT(PP2,S,T);
/*  THIS THE SOLUTION OF THE ADJOINT VARIABLES  */
LAM1(T);
LAM1(T):-"%;
LAM2(T);
LAM2(T):-"%;
LINSOLVE([LAM1(%PI/2)+1,LAM2(%PI/2)],[ALPHA,BETA]),GLOBALSOLVE:TRUE;
/* ALPHA BETA HAVE TO BE SUBSTITUTED IN THE ABOVE SOLUTION */
```

APPENDIX VI

```
/*  DESIGN SENSITIVITY ANALYSIS  */
DEPENDS([T,Z],[B]);
DEPENDS([ZZ],[T,B]);
DEPENDS([G],[ZZ,B]);
PSI:G+INTEGRATE(FF(Y,Z,B),Y,0,T);
DIFF(PSI,B);
EV(%,DIFF);
/*  NOTE: THE LAST COMMAND GIVES UNDESIRABLE RESULTS.
   (SEE REF.[ ]) THE PROBLEM CAN POSSIBLY BE HANDLED
   BY GRADEF [7] COMMAND   */
```

APPENDIX VII

```
/*........UNIT VECTORS ARE B1 B2 B3......SEE LEVINSON */
/*........DEFINE DOT AND CROSS PRODUCTS....*/
DOT(V1,V2):=BLOCK([P,PP],
FOR I:1 THRU 3 DO P[I]:RATCOEFF(V1,B[I]),
FOR I:1 THRU 3 DO PP[I]:RATCOEFF(V2,B[I]),
P[4]:SUM(P[I]*PP[I],I,1,3),
RETURN(P[4]))$
CROSS(V1,V2):=BLOCK([P,PP,PPP],
FOR I:1 THRU 3 DO P[I]:RATCOEFF(V1,B[I]),
FOR I:1 THRU 3 DO PP[I]:RATCOEFF(V2,B[I]),
PPP[1]:(P[2]*PP[3]-P[3]*PP[2]),
PPP[2]:(-P[1]*PP[3]+P[3]*PP[1]),
PPP[3]:(P[1]*PP[2]-P[2]*PP[1]),
PPP[4]:B[1]*PPP[1]+B[2]*PPP[2]+B[3]*PPP[3],
RETURN(PPP[4]))$
/*....NOW WE INPUT EQUATIONS FROM LEVINSON'S PAPER  */
DEPENDS(U,T);
DEPENDS(Q,T);
R[2]:COS(Q)*B[2]+SIN(Q)*B[3];
R[3]:-SIN(Q)*B[2]+COS(Q)*B[3];
WB:U[1]*B[1]+U[2]*B[2]+U[3]*B[3];
DERIVABBREV:TRUE;
U[4]:DIFF(Q,T);
WR:(U[1]+U[4])*B[1]+U[2]*B[2]+U[3]*B[3];
ALPB:DIFF(U[1],T)*B[1]+DIFF(U[2],T)*B[2]+DIFF(U[3],T)*B[3];
ALPR:DIFF(WR,T)+CROSS(WB,WR);
PPBS:B1*B[1]+B2*B[2]+B3*B[3];
```

```
PPRS:R1*B[1]+R2*R[2]+R3*R[3];
PRSBS:PPBS-PPRS;
VBS:U[5]*B[1]+U[6]*B[2]+U[7]*B[3];
VRS:VBS +DIFF(PRSBS,T)+CROSS(WB,PRSBS);
ABS:DIFF(VBS,T)+CROSS(WB,VBS);
ARS:DIFF(VRS,T)+CROSS(WB,VRS);
IBBSWB:BET1*B[1]*DOT(B[1],WB)+BET2*B[2]*DOT(B[2],WB)+BET3*B[3]*DOT(B[3],WB);
IRRSWR:RHO1*B[1]*DOT(B[1],WR)+RHO2*B[2]*DOT(B[2],WR)+RHO3*B[3]*DOT(B[3],WR);
IBBSALPB:BET1*B[1]*DOT(B[1],ALPB)+BET2*B[2]*DOT(B[2],ALPB)
      +BET3*B[3]*DOT(B[3],ALPB);
IRRSALPR:RHO1*B[1]*DOT(B[1],ALPR)+RHO2*B[2]*DOT(B[2],ALPR)
      +RHO3*B[3]*DOT(B[3],ALPR);
FSB:-MB*ABS;
FSR:-MR*ARS;
TSB:CROSS(IBBSWB,WB)-IBBSALPB;
TSR:CROSS(IRRSWR,WR)-IRRSALPR;
FB:F1*B[1]+F2*B[2]+F3*B[3];
TB:T1*B[1]+T2*B[2]+T3*B[3];
F[R]:=DOT(DIFF(VBS,U[R]),FB)+DOT(DIFF(WB,U[R]),TB);
FS[R]:=DOT(DIFF(VBS,U[R]),FSB)+DOT(DIFF(VRS,U[R]),FSR)
+DOT(DIFF(WB,U[R]),TSB)+DOT(DIFF(WR,U[R]),TSR);
EQ[R]:=F[R]+FS[R];
EQ[1];
FOR I:1 THRU 7 DO LDISPLAY ( X[1,I]:RATCOEFF(EQ[1],DIFF(U[I],T)));
```

APPENDIX VIII

```
XP:X-B*COS(TH);
YP:Y-B*SIN(TH);
XPP:X-A*COS(TH+GAM);
YPP:Y-A*SIN(TH+GAM);
EQ1:R**2-XP**2-YP**2;
EQ2:S**2-(XPP-P)**2-YPP**2;
EQ1:EXPAND(EQ1);
EQ1:RATSUBST(1-SIN(TH)^2,COS(TH)^2,%);
EQ1:EQ1/(2*B);
EQ2:EXPAND(EQ2);
EQ2:TRIGEXPAND(EQ2);
EQ2:RATSUBST(1-COS(TH)^2,SIN(TH)^2,EQ2);
EQ2:RATSUBST(1-COS(GAM)^2,SIN(GAM)^2,%);
EQ2:EQ2/(2*A);
LL:RATCOEF(EQ2,-SIN(TH));
NN:RATCOEF(EQ2,COS(TH));
MM:RATCOEF(EQ1,SIN(TH));
PP:RATCOEF(EQ1,-COS(TH));
PHPH:EQ1-MM*SIN(TH)+PP*COS(TH);
PSIPSI:EQ2+LL*SIN(TH)-NN*COS(TH);
PSIPSI:RATSIMP(%);

/* APPENDIX VIII (CONT.) */

EQ11:-P*COS(TH)+M*SIN(TH)-PH;
EQ22:N*COS(TH)-L*SIN(TH)-PSI;
SET:LINSOLVE([EQ11,EQ22],[COS(TH),SIN(TH)]);
CTH:RHS(PART(EV(SET),1));
STH:RHS(PART(EV(SET),2));
STH^2+CTH^2;
%-1;
XTHRU(%);
NUM(%);
```

APPENDIX IX

```
/*..........SNTHESIS BY ANALYTIC METHODS........*/
X2:A1*COS(PH);
Y2:A1*SIN(PH);
X3:-A4+A3*COS(PSI);
Y3:A3*SIN(PSI);
F1:A2^2-(X2-X3)^2-(Y2-Y3)^2;
EXPAND(%);
F11:%;
AA:RATCOEF(F11,SIN(PSI));
BB:RATCOEF(F11,COS(PSI));
F11-AA*SIN(PSI)-BB*COS(PSI);
RATSUBST(1-SIN(PH)^2,COS(PH)^2,%);
RATSUBST(1-SIN(PSI)^2,COS(PSI)^2,%);
T11:%;
T11:T11/(2*A1*A3);
/*.........SOLUTION BYDENAVIT METHOD...........*/
F1:A*SIN(PSI)+B*COS(PSI)-C;
RATSUBST(2*TAN(X)/(1+TAN(X)**2),SIN(PSI),%);
RATSUBST((1-TAN(X)**2)/(1+TAN(X)**2),COS(PSI),%);
RATSIMP(%);
SOLVE(%,TAN(X));

/*........SOLUTION BY HARTENBERG AND DENAVIT....*/
FF(PH,PSI):=K1*COS(PH)-K2*COS(PSI)+K3-COS(PH-PSI);
/*.............NOW A NUMERICAL EXAMPLE..........*/
/*........CHEBYCHEV SPACING IS GIVEN BY
XK=A+H*COS(2K-1)*PI/2N
    WHERE A=MEAN H=HALF THE INTERVAL OF X......*/
KEEPFLOAT:TRUE;
X(K):=3/2+1/2*COS((2*K-1)*%PI/6);
X1:X(3),NUMER;
X2:X(2);
X3:X(1),NUMER;
Y1:LOG(X1)/LOG(10),NUMER;
Y2:LOG(X2)/LOG(10),NUMER;
Y3:LOG(X3)/LOG(10),NUMER;
DELPH:60/180*%PI;
DELPSI:60/180*%PI;
PH1:0;
PSI1:0;
YF:LOG(2)/LOG(10),NUMER;

PH2:(X2-X1)*DELPH,NUMER;
PH3:(X3-X1)*DELPH,NUMER;
PSI2:(Y2-Y1)/YF*DELPSI,NUMER;
PSI3:(Y3-Y1)/YF*DELPSI,NUMER;
A4:1;
EQ1:FF(PH1,PSI1);
EQ2:FF(PH2,PSI2);
EQ3:FF(PH3,PSI3);
LINSOLVE([EQ1,EQ2,EQ3],[K1,K2,K3]),GLOBALSOLVE:TRUE;
A3:A4/K1;
A1:A4/K2;
A2:SQRT(A1**2+A3**2+A4**2-2*A1*A3*K3);
```

APPENDIX X

```
/*..ALGEBRA FOR QUATERNIONS FROM YANG'S PAPER..*/
NNPRED(N):=IS(N>=2);
NNPRED(2);
NNPRED(3);
MATCHDECLARE(NN,NNPRED);
TELLSIMPAFTER(EP^NN,0);
AL12H:AL12+EP*A12;
AL23H:AL23+EP*A23;
AL34H:AL34+EP*A34;
AL41H:AL41+EP*A41;
 TH1H:TH1+EP*S11;
 TH2H:TH2+EP*S2;
 TH3H:TH3+EP*S3;
 TH4H:TH4+EP*S4;
SAL12H: EXPAND(TAYLOR(SIN(AL12H),EP,0,1));
SAL23H: EXPAND(TAYLOR(SIN(AL23H),EP,0,1));
SAL34H: EXPAND(TAYLOR(SIN(AL34H),EP,0,1));
SAL41H: EXPAND(TAYLOR(SIN(AL41H),EP,0,1));
STH1H:  EXPAND(TAYLOR(SIN(TH1H),EP,0,1));
STH2H:  EXPAND(TAYLOR(SIN(TH2H),EP,0,1));
STH3H:  EXPAND(TAYLOR(SIN(TH3H),EP,0,1));
STH4H:  EXPAND(TAYLOR(SIN(TH4H),EP,0,1));
CAL12H:EXPAND(TAYLOR(COS(AL12H),EP,0,1));
CAL23H:EXPAND(TAYLOR(COS(AL23H),EP,0,1));
CAL34H:EXPAND(TAYLOR(COS(AL34H),EP,0,1));
CAL41H:EXPAND(TAYLOR(COS(AL41H),EP,0,1));
CTH1H: EXPAND(TAYLOR(COS(TH1H),EP,0,1));
CTH2H: EXPAND(TAYLOR(COS(TH2H),EP,0,1));
CTH3H: EXPAND(TAYLOR(COS(TH3H),EP,0,1));
CTH4H: EXPAND(TAYLOR(COS(TH4H),EP,0,1));
AATH1H:SAL12H*SAL34H*STH1H;
BBTH1H:-SAL34H*(SAL41H*CAL12H+CAL41H*SAL12H*CTH1H);
CCTH1H:CAL23H-CAL34H*(CAL41H*CAL12H-SAL41H*SAL12H*CTH1H);
EQ1:AATH1H*STH4H+BBTH1H*CTH4H-CCTH1H;
PRIMARY:EV(EQ1,EP=0);
DUAL:RATCOEFF(EQ1,EP);
A:RATCOEFF(PRIMARY,SIN(TH4));
B:RATCOEFF(PRIMARY,COS(TH4));
C:EXPAND(PRIMARY-A*SIN(TH4)-B*COS(TH4));
DUAL1:DUAL-S4*(A*COS(TH4)-B*SIN(TH4));
A0:RATCOEFF(DUAL1,SIN(TH4));
B0:RATCOEFF(DUAL1,COS(TH4));
CC0:EXPAND(DUAL1-A0*SIN(TH4)-B0*COS(TH4));
CC0:RATSIMP(CC0);
```

APPENDIX XI

```
/*....THIS PROGRAM SETS UP THE COMPLETE MATRIX
    EQATION FOR ROBOT....FOR DIFFERENTIAL MOTION....*/
/*....modified for Stanford manipulator....*/
/*CTH[I]:=COS(TH[I]);
STH[I]:=SIN(TH[I]); */
GCPRINT:FALSE;
CAL[I]:=COS(AL[I]);
SAL[I]:=SIN(AL[I]);
AL[1]:AL[4]:-%PI/2;
AL[2]:AL[5]:%PI/2;
AL[3]:AL[6]:0;
AA[1]:AA[3]:AA[4]:AA[5]:AA[6]:0;
AA[2]:0;
DD[1]:DD[4]:DD[5]:DD[6]:0;
CTH[3]:1;
STH[3]:0;
A[I]:=MATRIX([CTH[I],-STH[I]*CAL[I],STH[I]*SAL[I],AA[I]*CTH[I]],
[STH[I],CTH[I]*CAL[I],-CTH[I]*SAL[I],AA[I]*STH[I]],
[0,SAL[I],CAL[I],DD[I]],
[0,0,0,1]);
A[1];
RATSUBST(1-(STH[1])**2,(CTH[1])**2,%);
%^^-1;
IA[1]:%$
RATSUBST(1-(STH[1])**2,(CTH[1])**2,%);
IA[1]:%$
A[2];
%^^-1;
RATSUBST(1-(STH[2])**2,(CTH[2])**2,%);
IA[2]:%$
A[3];
%^^-1;
RATSUBST(1-(STH[3])**2,(CTH[3])**2,%);
IA[3]:%$
A[4];
%^^-1;
RATSUBST(1-(STH[4])**2,(CTH[4])**2,%);
IA[4]:%$
A[5];
%^^-1;
RATSUBST(1-(STH[5])**2,(CTH[5])**2,%);
IA[5]:%$
A[6];
%^^-1;
RATSUBST(1-(STH[6])**2,(CTH[6])**2,%);
IA[6]:%;
T56:A[6];
T46:A[5].A[6];
T36:A[4].A[5].A[6];
T26:A[3].A[4].A[5].A[6];
T16:A[2].A[3].A[4].A[5].A[6];
T6:A[1].A[2].A[3].A[4].A[5].A[6]$
NX:T6[1,1]$
NY:T6[2,1]$
NZ:T6[3,1]$
OX:T6[1,2]$
OY:T6[2,2]$
OZ:T6[2,3]$
AX:T6[1,3]$
AY:T6[2,3]$
AZ:T6[3,3]$
PX:T6[1,4]$
PY:T6[2,4]$
PZ:T6[3,4]$
TT6:MATRIX([NNX;OOX,AAX,PPX],
[NNY,OOY,AAY,PPY],
[NNZ,OOZ,AAZ,PPZ],
[0,0,0,1]);
IA1T6:IA[1].TT6;
IA2T6:IA[2].%;
IA3T6:IA[3].%;
IA4T6:IA[4].%$
IA5T6:IA[5].%$
IA6T6:IA[6].%$

/*....differential relations...this may be used
  for obtaining the sensitivity analysis...we
  follow the agorithm provided by R.P. Paul
  PAGE no.103....REVOLUTE=TDR PRISMATIC=TDP..*/

TDR(MAT):=BLOCK([NX,NY,NZ,PX,PY,PZ,OX,OY,OZ,AX,AY,AZ],
NX:MAT[1,1],NY:MAT[2,1],NZ:MAT[3,1],
OX:MAT[1,2],OY:MAT[2,2],OZ:MAT[3,2],
AX:MAT[1,3],AY:MAT[2,3],AZ:MAT[3,3],
PX:MAT[1,4],PY:MAT[2,4],PZ:MAT[3,4],
TRANSPOSE(MATRIX([-NX*PY+NY*PX,-OX*PY+OY*PX,-AX*PY+AY*PX,NZ,OZ,AZ])));
TDP(MAT):=BLOCK([NX,NY,NZ,PX,PY,PZ,OX,OY,OZ,AX,AY,AZ],
NX:MAT[1,1],NY:MAT[2,1],NZ:MAT[3,1],
OX:MAT[1,2],OY:MAT[2,2],OZ:MAT[3,2],
AX:MAT[1,3],AY:MAT[2,3],AZ:MAT[3,3],
PX:MAT[1,4],PY:MAT[2,4],PZ:MAT[3,4],
TRANSPOSE(MATRIX([NZ,OZ,AZ,0,0,0])));
/*....NOW WE SET UP COMPLETE DIFFERENTIAL MATRIX....*/
COL[1]:TDR(T6);
COL[2]:TDR(T16);
COL[3]:TDP(T26);
COL[4]:TDR(T36);
COL[5]:TDR(T46);
COL[6]:TDR(T56);
FOR J:1 THRU 6 DO (FOR I:1 THRU 6 DO (DIFFARRAY[I,J]:COL[J][I]));
JACOBIAN:GENMATRIX(DIFFARRAY,6,6);
/*....NUMERICAL EXAMPLE PAGE 107 ROBERT PAUL....*/
TH[1]:TH[4]:0;
TH[2]:TH[5]:TH[6]:%PI/2;
TH[3]:0;
DD[3]:20;

FOR I:1 THRU 6 DO (CTH[I]:COS(TH[I]),STH[I]:SIN(TH[I]));
JAC:EV(JACOBIAN,NUMER);
JAC:EV(%,NUMER);
DQ:TRANSPOSE(MATRIX([0.1,-0.1,2.0,0.1,0.1,0.1]));
JAC.DQ;
```

APPENDIX XII

```
/*  EXAMPLE OF ROBOT CONSIDERD BY LUMELSKY  */

CTH[I]:=COS(TH[I]);
STH[I]:=SIN(TH[I]);
GCPRINT:FALSE;
CAL[I]:=COS(AL[I]);
SAL[I]:=SIN(AL[I]);
AL[1]:AL[2]:AL[4]:%PI/2;
AL[3]:AL[6]:0;
AL[5]:-%PI/2;
AA[1]:AA[3]:AA[4]:AA[5]:AA[6]:0;
AA[2]:A2;
DD[1]:DD[2]:DD[4]:0;
CTH[3]:1;
STH[3]:0;
A[I]:=MATRIX([CTH[I],-STH[I]*CAL[I],STH[I]*SAL[I],AA[I]*CTH[I]],
[STH[I],CTH[I]*CAL[I],-CTH[I]*SAL[I],AA[I]*STH[I]],
[0,SAL[I],CAL[I],DD[I]],[0,0,0,1]);
A[1];
RATSUBST(1-STH[1]^2,CTH[1]^2,%);
%^^(-1);
A[1]:%$
RATSUBST(1-STH[1]^2,CTH[1]^2,%);
A[1]:%$
A[2];
%^^(-1);
RATSUBST(1-STH[2]^2,CTH[2]^2,%);
A[2]:%$
A[3];
%^^(-1);
RATSUBST(1-STH[3]^2,CTH[3]^2,%);
A[3]:%$
A[4];
%^^(-1);
RATSUBST(1-STH[4]^2,CTH[4]^2,%);
A[4]:%$
A[5];
%^^(-1);
RATSUBST(1-STH[5]^2,CTH[5]^2,%);
A[5]:%$
A[6];
%^^(-1);
RATSUBST(1-STH[6]^2,CTH[6]^2,%);
A[6]:%;
T56:A[6];
T46:A[5] . A[6];
T36:A[4] . A[5] . A[6];
T26:A[3] . A[4] . A[5] . A[6];
T16:A[2] . A[3] . A[4] . A[5] . A[6];
T6:A[1] . A[2] . A[3] . A[4] . A[5] . A[6]$
NX:T6[1,1];
NY:T6[2,1];
NZ:T6[3,1];
OX:T6[1,2];
```

```
OY:T6[2,2];
OZ:T6[2,3];
AX:T6[1,3];
AY:T6[2,3];
AZ:T6[3,3];
PX:T6[1,4];
PY:T6[2,4];
PZ:T6[3,4];
TT6:MATRIX([NNX,OOX,AAX,PPX],[NNY,OOY,AAY,PPY],[NNZ,OOZ,AAZ,PPZ],[0,0,0,1]);
A1T6:IA[1] . TT6;
A2T6:IA[2] . %;
A3T6:IA[3] . %;
A4T6:IA[4] . %;
A5T6:IA[5] . %;
A6T6:IA[6] . %;
EQ1:TT6-T6$
EQ2:IA1T6-T16$
EQ3:IA2T6-T26$
EQ4:IA3T6-T36$
EQ5:IA4T6-T46$
EQ6:IA5T6-T56$
TABLE(MAT,VAR):=BLOCK([EQ],FOR I THRU 4 DO (FOR J THRU 4 DO EQ[I,J]:0),
FOR I THRU 4 DO (FOR J THRU 4 DO (FOR L THRU 6 DO
(IF FREEOF(VAR[L],MAT[I,J]) = FALSE THEN EQ[I,J]:EQ[I,J]+T[L]
 ELSE FALSE))),GENMATRIX(EQ,4,4));
VAR[1]:TH[1];
VAR[2]:TH[2];
VAR[3]:DD[3];
VAR[4]:TH[4];
VAR[5]:TH[5];
VAR[6]:TH[6];
TABLE(EQ1,VAR);
TABLE(EQ2,VAR);
TABLE(EQ3,VAR);
TABLE(EQ4,VAR);
TABLE(EQ5,VAR);
TABLE(EQ6,VAR);
```

APPENDIX XII (CONT.)

FOLLOWING IS A PARTIAL OUTPUT FROM ABOVE PROGRAM:
T1 T2 ETC REPRESENTS PRESENCE OF VARIABLE 1 2 ETC.

TABLE(EQ2,VAR);

$$\text{COL 1} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_2 \\ T_6+T_5+T_4+T_1 \\ 0 \end{bmatrix} \quad \text{COL 2} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_2 \\ T_6+T_5+T_4+T_1 \\ 0 \end{bmatrix}$$

$$\text{COL 3} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_4+T_2 \\ T_5+T_4+T_1 \\ 0 \end{bmatrix} \quad \text{COL 4} = \begin{bmatrix} T_5+T_4+T_3+T_2+T_1 \\ T_5+T_4+T_3+T_2 \\ T_5+T_4+T_1 \\ 0 \end{bmatrix}$$

TABLE(EQ3,VAR);

$$\text{COL 1} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_1 \\ T_6+T_5+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 2} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_1 \\ T_6+T_5+T_2+T_1 \\ 0 \end{bmatrix}$$

$$\text{COL 3} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_4+T_1 \\ T_5+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 4} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_4+T_1 \\ T_5+T_3+T_2+T_1 \\ 0 \end{bmatrix}$$

TABLE(EQ4,VAR);

$$\text{COL 1} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_1 \\ T_6+T_5+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 2} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_4+T_1 \\ T_6+T_5+T_2+T_1 \\ 0 \end{bmatrix}$$

$$\text{COL 3} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_4+T_1 \\ T_5+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 4} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_4+T_1 \\ T_5+T_3+T_2+T_1 \\ 0 \end{bmatrix}$$

TABLE(EQ5,VAR);

$$\text{COL 1} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_2+T_1 \\ T_6+T_4+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 2} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_5+T_2+T_1 \\ T_6+T_4+T_2+T_1 \\ 0 \end{bmatrix}$$

$$\text{COL 3} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_2+T_1 \\ T_4+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 4} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_5+T_3+T_2+T_1 \\ T_4+T_2+T_1 \\ 0 \end{bmatrix}$$

TABLE(EQ6,VAR);

$$\text{COL 1} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_4+T_2+T_1 \\ T_5+T_4+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 2} = \begin{bmatrix} T_6+T_5+T_4+T_2+T_1 \\ T_6+T_4+T_2+T_1 \\ T_5+T_4+T_2+T_1 \\ 0 \end{bmatrix}$$

$$\text{COL 3} = \begin{bmatrix} T_5+T_4+T_2+T_1 \\ T_4+T_2+T_1 \\ T_5+T_4+T_2+T_1 \\ 0 \end{bmatrix} \quad \text{COL 4} = \begin{bmatrix} T_5+T_4+T_3+T_2+T_1 \\ T_4+T_2+T_1 \\ T_5+T_4+T_3+T_2+T_1 \\ 0 \end{bmatrix}$$

Appendix B

SOME REMARKS ON MACSYMA COMMANDS

We assume that the reader is familiar with the introduction given in the MACSYMA Primer, which is an introduction for beginners (see Reference [8]).

We will use *EXPR* to denote any symbolic expression such as $X+Y$, SIN(X), etc.

```
F: EXPR;
```

assigns *EXPR* to the variable F. (In the reference manual, F would be called an atomic variable.) Note that $F(X)$ has no meaning in this context. However, if we write

```
F(X):= EXPR
```

we are now defining a function $F(X)$. For example:

```
F(X):= SIN(X);
F(Z);                  (Machine prints SIN(Z))
```

This can also be achieved by the LAMBDA notation

```
F: LAMBDA([X],SIN(X));
```

We *can* now use $F(Z)$. For example:

```
F(Z);                  (Machine prints SIN(Z))
```

One advantage of this procedure is that F can be an argument to another function, for example as in SIMPSON in the text (Section 6).

An abbreviated, partial list of the commands that have been used in this report follows. For details, see the MACSYMA Manual.

DEPENDS([R,P],[RHO]);	$\equiv$ R and P are functions of ρ.
DEPENDENCIES(R(RHO));	$\equiv$ $R = R(\rho)$, as in the last command.
DIFF(SIN(X),X);	$\equiv$ $d\sin(x)/dx$
EV(EXPR,X=0);	$\equiv$ Evaluate *EXPR* with x=0. EV is a powerful command in MACSYMA and takes multiple arguments. See Manual.
GRADEF(R,RHO,P/R);	$\equiv$ set $\dfrac{\partial R}{\partial \rho} = P/R$
INTEGRATE(SIN(X),X);	$\equiv$ $\int \sin(x)\,dx$
LDISPLAY	$\equiv$ Display with equation numbers.
LINSOLVE([EQ1,EQ2],[X,Y]);	$\equiv$ Solve set of linear equations $Eq_1=0$, $Eq_2=0$, for x and y.
GLOBALSOLVE:TRUE;	$\equiv$ Assigns the values to the variables obtained by LINSOLVE command.
MAP(FACTOR,EXPR1);	$\equiv$ Factors each part of *EXPR* 1 separately.
RATCOEFF(EXPR,X^I)	$\equiv$ Obtain the coefficient of x^i in *EXPR*.

RATCOEFF(EXPR,X,I) ≡ Same as above. If the third argument is not specified, as in the last command, it is taken as 1 by default.

RATSIMP(EXPR); ≡ Obtain rational simplification of expression *EXPR*.

SUBST(0,X,EXPR); ≡ Substitute 0 for x in the expression *EXPR*.

RATSUBST(0,X,EXPR1); ≡ Same as SUBST after expansion.

SUM(P[I]X^I,I,0,M); ≡ $\sum_{i=0}^{M} P_i x^i$

TAYLOR(SIN(X),X,0,5); ≡ Obtain Taylor expansion of $\sin x$ around $x=0$ up to the fifth power. Can also be used for multivariate functions.

TRIGEXPAND(EXPR1); ≡ Expand the trigonometric functions in the expression *EXPR*1.

FOR I:1 STEP 1 THRU N DO (ANY MACSYMA COMMAND); (Involving set of *I*'s)
≡ This is a DO loop for 1 to n in steps of 1. Note that the default for starting is 1 and that the default step is 1; FOR I THRU N DO(...) will accomplish the same task.

ON THE CONVERSION OF LINPACK TO ADA

Benjamin J. Martin
Department of Mathematical and Computer Sciences
Atlanta University
Atlanta, Georgia 30314

Robert Bozeman
Department of Mathematics
Morehouse College
Atlanta, Georgia 30314

ABSTRACT. The authors demonstrate the feasibility of converting the LINPACK routines for analyzing and solving systems of linear algebraic equations from FORTRAN to Ada. This is done with minimal alteration of the original program structure, thus requiring very little re-orientation by current users of LINPACK. Sample programs are included.

I. INTRODUCTION. In a paper entitled "Can Ada Replace FORTRAN for Numerical Computation?", Alfred Morris, Jr., of Naval Surface Weapons Center in Dahlgren, Virginia, argues that Ada is an inadequate substitute for FORTRAN. He presents several points which he considers to be critical deficiencies of Ada. Among these "critical deficiencies" is the failure of the Ada specifications to include the internal representation of arrays. The FORTRAN standard requires that arrays be stored in column major form, that is, columns are stored together one after the other. This standard along with the absence of strong typing allows the programmer to access the elements of a matrix as if it were a vector and always get the right element. It is this standard, among others, that has permitted the development of the LINPACK routines.

Mr. Morris points out that this problem leads to the necessity of abandoning a large selection of fully developed algorithms and routines. Says he, "it is clear that a considerable portion of a quarter of century accumulation of logic and code could not be adapted to Ada. This would include software packages such as Argonne National Laboratories' LINPACK ... which is highly refined and quite widely used on a variety of computers." It is the aim of this paper to demonstrate the feasibility of recoding the LINPACK routines in the Ada language.

II. METHODS OF CONVERSION. There are three approaches to the recoding of the LINPACK routines which seem obvious and straightforward. The first approach is to merely insert the appropriate codes for the various BLAS routines in the various places where the subroutine calls are made. This would reduce the work required in performing the conversion. The effect of this insertion on execution time should be minimal since no extra code is being executed. In fact, the overhead of the subroutine call is saved. However, it would significantly increase the space requirements for the program code. It would also destroy the modularity and the readibility of the routines.

The second approach is to process the matrices and vectors prior to the BLAS subroutine call and postprocess them after the BLAS subroutine call. The preprocessor would remove the appropriate portion of the column of the matrix or the appropriate portion of the vector. This approach is placed back in its place in the matrix or vector. This approach maintains the readability and modularity of the program. The increase in space requirements is minimal in that only four short routines are needed in addition to the usual ones. The increase in time requirements is also minimal since the only thing these routines do is transfer several data items back and forth.

The third approach is to define the matrix to be an array of vectors. After the matrix to be used is properly defined the appropriate vector is sent to the BLAS subroutine. In order to use this technique the vectors to be transferred must be the columns of the matrix. This may require a routine to compute the transpose of the matrix before proceeding. Therefore, the increase in space requirements and in execution time should be minimal. The Ada concept of slices may prove useful in this approach.

III. THE IMPLEMENTATION. The first alternative was dismissed as being the least attractive alternative. The second alternative appeared to be the easiest to implement quickly, and so was considered first. The third alternative is presently being pursued.

In order to implement the second alternatives, LINPACK and BLAS routines for a general system were coded in Ada. The changes made in the programs were of two types. The first type of changes simply involved the use of structured programming technique making use of control structures not available in FORTRAN. The second type of change involved writing four new routines called CONVRTM, RECONVRTM, CONVRTV, and RECONVRTV. The first two routines work on matrices. CONVRTM takes a given portion of a column from a matrix and stores it in a vector. RECONVRTM takes a portion of a column from a vector and replaces it in a matrix. The other two routines do similar things for a vector. The routines are used in conjunction with the BLAS routines as follows:

```
The FORTRAN statement   CALL SAXPY(N-K,T,A,(K+1,K),1,B(K+1),1)
is replaced by the Ada sequence   CONVRTM(N-K,K+1,K,A,X,1);
                                  CONVRTV(N-K,K+1,B,Y);
                                  RECONVRTM(N-K,K+1,K,A,X,1);
                                  RECONVRTV(N-K,K+1,B,Y);
```

Attached is the program which resulted from this implementation and which clearly defines the action of each of these routines. These routines were coded under ADAED, version 16.3. They may or may not run under other versions of ADAED. Because of the nature of ADAED, extensive testing is not possible. A simple system of five equations in five unknowns took 30 minutes of CPU time to compile and execute.

IV. CONCLUSIONS. While it may not be possible to retain all of the characteristics of the LINPACK routines in a conversion to Ada, it has been demonstrated that such a conversion is possible. Because of the limitations of ADAED, especially its execution speed, any real testing must await a compiler. Nevertheless, it is clear that the routines can be fairly easily recoded in Ada. The costs incurred in this recoding cannot be determined at this time. Other approaches may also be available besides the ones indicated above. The third alternative may be the nest of the three. There are indeed unanswered questions, but we must conclude that it is feasible to salvage at least a portion of a "quarter century accumulation of logic and code."

Bibliography

Dongarra, J.J., *LINPACK User's Guide*, Siam Press, Philadelphia, PA., 1979.

Morris, A.H.,Jr., "Can Ada Replace FORTRAN for Numerical Computation?" *ACM, Sigplan Notices*, vol.16, number 12,(Dec. 1981).

```
WITH TEXT_IO;
PROCEDURE TEST IS
USE TEXT_IO;
TYPE REAL IS NEW FLOAT;
PACKAGE HIS_IO IS NEW INTEGER_IO(INTEGER);
PACKAGE MY_IO IS NEW FLOAT_IO(REAL);
USE HIS_IO; USE MY_IO;
SUBTYPE INDEX IS INTEGER RANGE 1..5;
TYPE INTVEC IS ARRAY (INDEX) OF INDEX;
TYPE VECTOR IS ARRAY (INDEX) OF REAL ;
TYPE MATRIX IS ARRAY (INDEX,INDEX) OF REAL;
   A:MATRIX;
   B:VECTOR;
   IPVT:INTVEC;
   INFO,JOB:INTEGER:=0;
   SIZE:INDEX:=5;
--
FUNCTION ISAMAX(N:INDEX;X:VECTOR) RETURN INDEX IS
--
--FINDS THE INDEX OF ELEMENT HAVING MAXIMUM ABSOLUTE VALUE.
--ADAPTED FROM THE 3/11/78 VERSION BY JACK DONGARRA
--
   IMAX:INDEX:=1;
   XMAX:REAL;
BEGIN
   IF N<1 THEN RETURN 0; END IF;
   IF N=1 THEN RETURN 1; END IF;
   XMAX:=ABS(X(1));
       FOR I IN 2..N LOOP
            IF ABS(X(I))>XMAX THEN
                    IMAX:=I;
                    XMAX:=ABS(X(I));
            END IF;
       END LOOP;
   RETURN IMAX;
END ISAMAX;
--
--
FUNCTION SASUM(N:INDEX;X:VECTOR) RETURN REAL IS
--
--TAKES THE SUM OF THE ABSOLUTE VALUES.
--USES UNROLLED LOOPS.
--ADAPTED FROM THE 3/11/78 VERSION BY JACK DONGARRA
--
M2:INTEGER;
M,M1:INDEX;
SUM:REAL:=0.0;
BEGIN
  IF N=0 THEN RETURN SUM; END IF;
  M2:=N MOD 6;
  IF M2 /= 0 THEN
          M:=M2;
          FOR I IN 1..M LOOP
              SUM:=SUM+ABS(X(I));
          END LOOP;
  END IF;
  IF N >= 6 THEN
        IF M2=0 THEN M1:=1; ELSE M1:=M+1;END IF;
        WHILE M1<N LOOP
           SUM:=SUM+ABS(X(M1))+ABS(X(M1+1))+ABS(X(M1+2))+ABS(X(M1+3))
```

```
                +ABS(X(M1+4))+ABS(X(M1+5));
           M2:=M1+6;
           IF M2<N THEN M1:=M2; ELSE M1:=N; END IF;
        END LOOP;
  END IF;
  RETURN SUM;
END SASUM;
--
--
PROCEDURE SAXPY(N:INDEX;S:REAL;X:VECTOR;Y: IN OUT VECTOR) IS
--
--COMPUTES A CONSTANT TIMES A VECTOR PLUS A VECTOR.
--USES UNROLLED LOOPS.
--ADAPTED FROM THE 3/11/78 VERSION BY JACK DONGARRA
--
M,M1:INDEX;
M2:INTEGER;
BEGIN
  IF N <= 0 THEN RETURN; END IF;
  IF S=0.0 THEN RETURN; END IF;
  M2:=N MOD 4;
  IF M2 /= 0 THEN
        M:=M2;
        FOR I IN 1..M LOOP
           Y(I):=Y(I)+S*X(I);
        END LOOP;
  END IF;
  IF N >= 4 THEN
        IF M2=0 THEN M1:=1; ELSE M1:=M+1; END IF;
        WHILE M1<N LOOP
          Y(M1):=Y(M1)+S*X(M1);
          Y(M1+1):=Y(M1+1)+S*X(M1+1);
          Y(M1+2):=Y(M1+2)+S*X(M1+2);
          Y(M1+3):=Y(M1+3)+S*X(M1+3);
          M2:=M1+4;
          IF M2<N THEN M1:=M2; ELSE M1:=N; END IF;
        END LOOP;
  END IF;
EXCEPTION
  WHEN CONSTRAINT_ERROR=>
  PUT ("CONSTRAINT ERROR IN SAXPY N, M1, M: "); PUT(N);
  PUT(M1);PUT(M);
END SAXPY;
--
--
FUNCTION SDOT(N:INDEX;X,Y:VECTOR) RETURN REAL IS
--
--FORMS THE DOT PRODUCT OF TWO VECTORS.
--USES UNROLLED LOOPS.
--ADAPTED FROM THE 3/11/78 VERSION BY JACK DONGARRA
--
M2:INTEGER;
M,M1:INDEX;
TEMP:REAL:=0.0;
BEGIN
   IF N<=0 THEN RETURN 0.0; END IF;
   M2:=N MOD 5;
   IF M2 /= 0 THEN
      M:=M2;
      FOR I IN 1..M LOOP
        TEMP:=TEMP+X(I)*Y(I);
      END LOOP;
   END IF;
   IF N>=5 THEN
      IF M2=0 THEN M1:=1; ELSE M1:=M+1;END IF;
      WHILE M1<N LOOP
        TEMP:=TEMP+X(M1)*Y(M1);
        M2:=M1+5;
        IF M2<N THEN M1:=M2; ELSE M1:=N; END IF;
      END LOOP;
   END IF;
   RETURN TEMP;
END SDOT;
--
--
PROCEDURE SSCAL(N:INDEX;S:REAL;X:IN OUT VECTOR) IS
--
--SCALES A VECTOR BY A CONSTANT.
--USES UNROLLED LOOPS
--ADAPTED FROM THE 3/11/78 VERSION BY JACK DONGARRA
--
M2:INTEGER;
M,M1:INDEX;
BEGIN
   IF N<=0 THEN RETURN; END IF;
   M2:=N MOD 5;
   IF M2 /= 0 THEN
     M:=M2;
     FOR I IN 1..M LOOP
       X(I):=S*X(I);
     END LOOP;
   END IF;
   IF N>=5 THEN
     IF M2=0 THEN M1:=1; ELSE M1:=M+1;END IF;
     WHILE M1<N LOOP
        X(M1):=S*X(M1);
        X(M1+1):=S*X(M1+1);
        X(M1+2):=S*X(M1+2);
        X(M1+3):=S*X(M1+3);
        X(M1+4):=S*X(M1+4);
        M2:=M1+5;
        IF M2<N THEN M1:=M2; ELSE M1:=N; END IF;
     END LOOP;
   END IF;
   RETURN;
END SSCAL;
--
PROCEDURE CONVRTM(N,K,L:INDEX;A:MATRIX;V:OUT VECTOR; INC:INDEX) IS
BEGIN
        IF INC=1 THEN                  -- PLACE COLUMN IN VECTOR
                FOR I IN 1..N LOOP
                        V(I):=A(K+I-1,L);
                END LOOP;
        ELSIF INC>1 THEN               -- PLACE ROW IN VECTOR
                FOR I IN 1..N LOOP
                        V(I):=A(K,L+I-1);
                END LOOP;
        END IF;
END CONVRTM;
--
--
PROCEDURE CONVRTV(N,K:INDEX;B:VECTOR;V: OUT VECTOR) IS
```

```
BEGIN
        FOR I IN 1..N LOOP
                V(I):=B(K+I-1);
        END LOOP;
END CONVRTV;
PROCEDURE RECONVRTM(N,K,L:INDEX;A:OUT MATRIX;V:VECTOR;INC:INDEX) IS
BEGIN
        IF INC=1 THEN                   -- PLACE COLUMN IN VECTOR
                FOR I IN 1..N LOOP
                        A(K+I-1,L):=V(I);
                END LOOP;
        ELSIF INC>1 THEN               -- PLACE ROW IN VECTOR
                FOR I IN 1..N LOOP
                        A(K,L+I-1):=V(I);
                END LOOP;
        END IF;
END RECONVRTM;
--
--
PROCEDURE RECONVRTV(N,K:INDEX;B:OUT VECTOR;V:VECTOR) IS
BEGIN
        FOR I IN 1..N LOOP
                B(K+I-1):=V(I);
        END LOOP;
END RECONVRTV;
PROCEDURE SGESL(A:IN OUT MATRIX;LDA,N:INDEX;IPVT:IN OUT INTVEC;
                                        B:IN OUT VECTOR;JOB:INTEGER) IS

--
--      SGESL SOLVES THE REAL SYSTEM   A * X = B
--                              OR     TRANS(A) * X = B
--      USING THE FACTORS COMPUTED BY SGECO OR SGEFA.
--
--      ON ENTRY
--
--        A     THE OUTPUT FROM SGECO OR SGEFA.
--
--        LDA   INDEX
--              THE LEADING DIMENSION OF THE ARRAY A.
--
--        N     INDEX
--              THE ORDER OF THE MATRIX A.
--
--        IPVT  THE INDEX PIVOT VECTOR FROM SGECO OR SGEFA.
--
--        B     THE REAL RIGHT HAND SIDE VECTOR.
--
--        JOB   INDEX
--              = 0     TO SOLVE A*X=B.
--              <>0     TO SOLVE TRANS(A)*X=B WHERE TRANS(A) IS THE TRANSPOSE.
--
--      ON EXIT
--
--        B     THE SOLUTION VECTOR X.
--
--      ERROR CONDITION
--
--        A DIVISION BY ZERO WILL OCCUR IF THE INPUT FACTOR CONTAINS A ZERO ON
--        THE DIAGONAL.  TECHNICALLY THIS INDICATES SINGULARITY BUT IT IS OFTEN
--        CAUSED BY IMPROPER ARGUMENTS OR IMPROPER SETTING OF LDA.  IT WILL NOT
--        OCCUR IF THE SUBROUTINES ARE CALLED CORRECTLY AND IF SGECO HAS SET
--          RCOND>0.0 OR SGEFA HAS SET INFO=0.
--
--        TO COMPUTE INVERSE(A)*C WHERE C IS A MATRIX WITH P COLUMNS
--                SGECO(A,LDA,N,IPVT,RCOND,Z);
--                IF (RCOND NOT IS TOO SMALL) THEN
--                   FOR J IN 1..P LOOP
--                      SGESL(A,LDA,N,IPVT,C(1,J));
--                   END LOOP;
--                END IF;
--
--        LINPACK.  THIS VERSION IS BASED ON THE 08/14/78 VERSION BY
--        CLEVE MOLER, UNIVERSITY OF NEW MEXICO, ARGONNE NATIONAL LAB.
--
--        PROCEDURES AND FUNCTIONS:  BLAS, SAXPY, SDOT
--
K,KB,L,NM1:INDEX;    --INTERNAL VARIABLES
T: REAL;
X,Y:VECTOR;
BEGIN
     NM1:=N-1;
     IF JOB=0 THEN                       -- SOLVE A*X=B
        IF NM1>0 THEN                    -- FIRST SOLVE L*Y=B
             FOR K IN 1..NM1 LOOP
                L:=IPVT(K);
                T:=B(L);
                IF L/=K THEN
                     B(L):=B(K);
                     B(K):=T;
                END IF;
                CONVRTM(N-K,K+1,K,A,X,1);
                CONVRTV(N-K,K+1,B,Y);
                SAXPY(N-K,T,X,Y);
                RECONVRTM(N-K,K+1,K,A,X,1);
                RECONVRTV(N-K,K+1,B,Y);
             END LOOP;
        END IF;
        FOR KB IN 1..N-1 LOOP         -- NOW SOLVE U*X=Y
             K:=N+1-KB;
             B(K):=B(K)/A(K,K);
             T:=-B(K);
             CONVRTM(K-1,1,K,A,X,1);
             CONVRTV(K-1,1,B,Y);
             SAXPY(K-1,T,X,Y);
             RECONVRTM(K-1,1,K,A,X,1);
             RECONVRTV(K-1,1,B,Y);
        END LOOP;
        B(1):=B(1)/A(1,1);     --THIS AVOIDS A CALL TO SAXPY WITH K-1=0
     ELSE            -- JOB<>0, SOLVE TRANS(A)*X=B
        B(1):=B(1)/A(1,1);        -- THIS AVOIDS A CALL TO SDOT WITH K-1=0
        FOR K IN 2..N LOOP    -- FIRST SOLVE TRANS(U)*Y=B
             CONVRTM(K-1,1,K,A,X,1);
             CONVRTV(K-1,1,B,Y);
             T:=SDOT(K-1,X,B);
             B(K):=(B(K)-T)/A(K,K);
        END LOOP;
        IF NM1>0 THEN         -- NOW SOLVE TRANS(L)*X=Y
             FOR KB IN 1..NM1 LOOP
                K:=N-KB;
                CONVRTM(N-K,K+1,K,A,X,1);
                CONVRTV(N-K,K+1,B,Y);
                B(K):=B(K)+SDOT(N-K,X,Y);
                L:=IPVT(K);
                IF L/=K THEN
                     T:=B(L);
                     B(L):=B(K);
                     B(K):=T;
                END IF;
             END LOOP;
        END IF;
     END IF;
END SGESL;
PROCEDURE SGEFA(A:IN OUT MATRIX;LDA,N:INDEX;IPVT:IN OUT INTVEC;
                                                INFO: OUT INTEGER) IS
--
--      SGEFA FACTORS A PEAL MATRIX BY GAUSSIAN ELIMINATION
--
--      SGEFA IS USUALLY CALLED BY SGECO, BUT IT CAN BE CALLED DIRECTLY WITH
--      A SAVING IN TIME IF RCOND IS NOT NEEDED.
--      (TIME FOR SEGCO) = (1 + 9/N)*(TIME FOR SGEFA).
--
--      ON ENTRY
--
--         A    THE REAL MATRIX TO BE FACTORED.
--
--         LDA  INDEX
--              THE LEADING DIMENSION OF THE ARRAY A.
--
--         N    INDEX
--              THE ORDER OF THE MATRIX A.
--
--      ON RETURN
--
--         A    AN UPPER TRIANGULAR MATRIX AND THE MULTIPLIERS WHICH WERE USED
--              TO OBTAIN IT.  THE FACTORIZATION CAN BE WRITTEN A = L*U WHERE
--              L IS A PRODUCT OF PERMUTATINS AND UNIT LOWER TRIANGULAR MATRICE
--              AND U IS UPPER TRIANGULAR.
--
--         IPVT AN INDEX VECTOR OF PIVOT INDICES.
--
--         INFO INTEGER
--              =0  NORMAL VALUE.
--              =K  IF U(K,K) = 0.  THIS IS NOT AN ERROR CONDITION FOR THIS
--                  SUBROUTINE, BUT IT DOES INDICATE THAT SGESL OR SGEDI WILL
--                  DIVIDE BY ZERO IF CALLED.  USE RCOND IN SGECO FOR A RELIABL
--                  INDICATION OF SINGULARITY.
--
--      LINPACK.  THIS VERSION BASED ON THE 08/14/78 VERSION BY
--      CLEVE MOLER, UNIVERSITY OF NEW MEXICO, ARGONNE NATIONAL LAB.
--
--      SUBROUTINES AND FUNCTIONS: BLAS. SAXPY, SSCAL, ISAMAX
--
J,K,KP1,L,NM1:INDEX;  --  INTERNAL VARIABLES
T:REAL;
X,Y:VECTOR;
BEGIN
--
--      GAUSSIAN ELIMINATION WITH PARTIAL PIVOTING
--
        INFO:=0;
        NM1:=N-1;
        IF NM1>0 THEN
            FOR K IN 1..NM1 LOOP
```

```
                KP1:=K+1;
--
--      PULL OUT THE APPROPRIATE COLUMN VECTOR FOR ISAMAX
--
            CONVRTM(N-K+1,K,K,A,X,1);
            L:=ISAMAX(N-K+1,X)+K-1;      -- FIND L = PIVOT INDEX
            IPVT(K):=L;    -- ZERO PIVOT IMPLIES THIS COLUMN ALREADY
                IF A(L,K)=0.0E0 THEN INFO:=INTEGER(K);           -- TRIANGULAR
                ELSE
                      IF L/=K THEN       --      INTERCHANGE IF NECESSARY
                          T:=A(L,K);
                          A(L,K):=A(K,K);
                          A(K,K):=T;
                      END IF;
                      T:=-1.0E0/A(K,K);   -- COMPUTE MULTIPLIERS
                      CONVRTM(N-K,K+1,K,A,X,1);
                      SSCAL(N-K,T,X);
                      RECONVRTM(N-K,K+1,K,A,X,1);
                      FOR J IN KP1..N LOOP  --  ROW ELIMINATION WITH
                          T:=A(L,J);                     -- COLUMN INDEXING
                          IF L/=K THEN
                                A(L,J):=A(K,J);
                                A(K,J):=T;
                          END IF;
                          CONVRTM(N-K,K+1,K,A,X,1);
                          CONVRTM(N-K,K+1,J,A,Y,1);
                          SAXPY(N-K,T,X,Y);
                          RECONVRTM(N-K,K+1,K,A,X,1);
                          RECONVRTM(N-K,K+1,J,A,Y,1);
                      END LOOP;
                END IF;
            END LOOP;
        END IF;
        IPVT(N):=N;
        IF A(N,N)=0.0E0 THEN INFO:=N; END IF;
END SGEFA;
BEGIN
PUT_LINE("TYPE IN THE MATRIX A");
   FOR I IN INDEX LOOP
       FOR J IN INDEX LOOP
          GET(A(I,J));
       END LOOP;
       GET(B(I));
   END LOOP;
   FOR I IN INDEX LOOP
     FOR J IN INDEX LOOP
       PUT(A(I,J));PUT("  ");
     END LOOP;
       NEW_LINE;
       PUT(B(I));
       NEW_LINE;
END LOOP;
   SGEFA(A,SIZE,SIZE,IPVT,INFO);
   SGESL(A,SIZE,SIZE,IPVT,B,JOB);
   FOR I IN INDEX LOOP
     PUT(B(I));PUT("  ");
   END LOOP;
   NEW_LINE;
END TEST;
```

AUTOMATIC GENERATION OF TAYLOR SERIES IN PASCAL-SC: BASIC APPLICATIONS TO ORDINARY DIFFERENTIAL EQUATIONS

George Corliss
Department of Mathematics, Statistics, and Computer Science
Marquette University, Milwaukee WI 53233

and

L. B. Rall*
Mathematics Research Center
University of Wisconsin-Madison, Madison WI 53706

ABSTRACT. Taylor series have a long history of usefulness in numerical analysis, especially for the numerical solution of the initial-value problem for systems of ordinary differential equations. Recurrence relations for coefficients of Taylor series are well-known, for arithmetic operations and various standard functions with series arguments. Compilers for languages such as Pascal-SC, Algol 68, and Ada™ (a trademark of the U. S. Department of Defense) have built-in facilities for the support of user-defined data types and operators which allow automatic generation of machine code to evaluate Taylor coefficients. In addition, Pascal-SC (Pascal for Scientific Computation) offers accurate floating-point and interval arithmetic for numerical calculations. In this language, series with real (interval) coefficients are introduced as type TAYLOR (ITAYLOR). The operators +, -, *, /, ** and the functions SQR, SQRT, EXP, SIN, COS, ARCTAN, and LN are implemented for arguments of types TAYLOR, ITAYLOR, INTEGER, REAL, and INTERVAL. An initial-value problem for an ordinary differential equation is solved using types TAYLOR and ITAYLOR. A stability analysis shows that the recurrence relations for the series generation exhibit a mild instability which has no significant effect on the values of the solution computed by analytic continuation.

AMS (MOS) Subject Classifications: 65-04, 65G10, 65L05, 65L07, 65V05

Key Words and Phrases: Taylor series, recurrence relations for Taylor coefficients, automatic differentiation, numerical solution of ordinary differential equations, stability, error analysis, interval arithmetic

1. TAYLOR SERIES, POLYNOMIALS, AND FORMS.

A fundamental tool of numerical analysis is the expansion of a real function f of a real variable x into a Taylor series at $x = x_0$, which gives the expression

$$f(x) = \sum_{i=1}^{\infty} f^{(i-1)}(x_0)(x - x_0)^{(i-1)}/(i-1)!, \tag{1.1}$$

valid for $|x - x_0| < \rho$, where ρ is the radius of convergence of the infinite series on the right-hand side of (1.1). Of course, in actual numerical computation, the Taylor polynomial

*Research sponsored by the U. S. Army under Contract No. DAAG29-80-C-0041.

(1.2) $$f_n(x) = \sum_{i=1}^{n} f^{(i-1)}(x_0)(x - x_0)^{(i-1)}/(i-1)!,$$

is used in place of the infinite series. This results in the truncation error

(1.3) $$R_n(f,x;x_0) = f(x) - f_n(x) = f^{(n)}(\xi)(x - x_0)^n/n!, \quad \xi \in X,$$

where X denotes the interval $X = [\min\{x,x_0\},\max\{x,x_0\}]$, and the remainder term $R_n(f,x_0;x)$ is expressed in Lagrange form. This approximation of f(x) by $f_n(x)$ gives rise to a problem of error estimation which can be solved by the methods of interval analysis. If $F^{(n)}$ is an interval inclusion of the real function $f^{(n)}$, then

(1.4) $$f(x) - f_n(x) = R_n(f,x_0;x) \in F^{(n)}(X)(x - x_0)^n/n!;$$

this allows automatic computation of guaranteed error bounds by the use of interval arithmetic [13], [14].

In order for Taylor series methods to be useful in scientific computation, it must be possible to automate the calculation of the normalized real Taylor coefficients

(1.5) $$c(i + 1) = f^{(i)}(x_0)(x - x_0)^i/i!, \quad i = 1,\ldots,n-1,$$

and the corresponding interval quantities

(1.6) $$C(i + 1) = F^{(i)}(X)(x - x_0)^i/i!, \quad i = 1,\ldots,n-1.$$

These calculations can be carried out by means of well-known recurrence relations [1], [13], [14], [17] for functions defined by subroutines or expressions involving arithmetic operations and a variety of standard functions for which library subroutines are available. A very important application of automated generation of Taylor series by recursion is the numerical solution of the initial-value problem for ordinary differential equations. That is, it is required to find $y = y(x) = (y_1(x),y_2(x),\ldots,y_m(x))$ such that

(1.7) $$y_i' = f_i(x,y), \quad y_i(x_0) = y_{i0}, \quad i = 1,\ldots,m,$$

for values of x in an interval containing x_0 [1], [3], [6], [13], [14].

Another application of the methods in this paper is to the automatic generation of interval inclusions of real functions by means of their interval mean-value and Taylor forms [13], [14], [20]. Suppose, for example, that f(x) is a real function, such as

(1.8) $$f(x) = (x + 3)/(x^2 + 2),$$

which can be evaluated by the corresponding expression

(1.9) `f := (x + 3)/(x**2 + 2);`

in a Pascal-SC program. An interval inclusion F of f on an interval X, for which

(1.10) $$f(X) = \{f(x) \mid x \in X\} \subset F(X)$$

can be obtained simply by declaring the variables F and X to be of type INTERVAL, and then evaluating the expression corresponding to (1.9),

$$F := (X + 3)/(X**2 + 2) \tag{1.11}$$

using interval arithmetic, a standard feature of Pascal-SC [24]. An inclusion obtained in this way may be too coarse in the sense that F(X) is a much larger interval than needed to contain f(X). In this case, an interval inclusion provided by the mean-value form

$$F_1(X) = f(x) + F'(X)(X - x), \quad x \in X, \tag{1.12}$$

can be better, particularly if the width of X is not large [13], [14], [20]. In (1.12), F' denotes an interval inclusion of the derivative f' of f; F'(X) is obtained automatically by evaluating (1.11) with F and X of type ITAYLOR, as will be explained below. Interval inclusions of f are also provided by Taylor forms of higher order [20], in general,

$$F_n(X) = \sum_{i=0}^{n-1} f^{(i)}(x)(X - x)^i/i! + F^{(n)}(X)(X - x)^n/n!, \quad x \in X. \tag{1.13}$$

These forms can be generated automatically from the expressions (1.9) and (1.11) by the use of types TAYLOR and ITAYLOR, respectively. Recursive generation of real and interval Taylor coefficients makes possible an adaptive method for calculation of interval inclusions of real functions, in which n is increased until $F_n(X)$ includes $F_{n-1}(X)$. It is also possible to reduce the width of computed inclusions by making use of the fact that the intersection of interval inclusions is likewise an interval inclusion.

Previous implementations of automatic generation of Taylor coefficients in computer languages such as FORTRAN have used interpretation [21] or pre-compilation [11] to activate the necessary subroutines [17]. In more modern languages, the compiler itself can be used to produce the necessary routines, leading to a saving of programming effort and an increase in clarity of the source code. The use of Pascal-SC, a language of this type, will be explained in the next section.

2. PASCAL-SC.

The method for automatic generation of Taylor series given in this report is based on computation with the coefficients of Taylor polynomials of arbitrary length, considered as specific mathematical entities. This requires that the language support i) user defined data types, as do descendents of ALGOL-60 such as Pascal and ADA™ (ADA is a trademark of the U.S. Department of Defense); and ii) user defined operators, as do ALGOL-68 and ADA.

Pascal-SC [2] is an extension of Pascal which provides both user-defined data types and user-defined operators. This paper assumes a modest familiarity with standard Pascal [9]. For the remainder of this Section, we outline some of the extensions which make Pascal-SC well suited to the applications in this paper. The reader who wishes to omit the discussion of programming language issues may proceed directly to the definition of the data types TAYLOR and ITAYLOR in Section 3.

Pascal-SC was developed with the needs of scientific computation in mind. It is an implementation of Jensen and Wirth Pascal [9] which also provides intervals, complex numbers, complex intervals, as built-in elementary scalar data types [24]. A full range of standard operators is provided to manipulate the elementary scalar data types, as well as vectors and matrices built of these types [24].

Standard Pascal supports user-defined data types built from elementary data types. This feature will be used to define variables of type TAYLOR and ITAYLOR (interval Taylor) in Section 3.

Pascal-SC allows the user to define operators. Most computer languages allow programmers to define functions, subroutines, or procedures, but except for APL, the languages most often used for scientific computation require that such user-defined functions be called using a prefix notation (eg. SIN (X)), while built-in operators are called using an infix notation (eg. A + B). Programmers can define operators in Pascal-SC to extend the language in a uniform way, retaining the familiar infix notation for operators whose operands are variables of user-defined types (eg. A + B, where A and B are variables of type TAYLOR).

Operators, functions, and procedures in Pascal-SC can be overloaded. That is, the name of an operator, a function, or a procedure can have different meanings, depending on the type or number of its operands. For example, the standard Pascal or FORTRAN operator "+" is said to be overloaded because "A + B" for integer variables A and B has a different meaning from "A + B" for real variables A and B. The support of Pascal-SC for overloading of user defined operators is essential to the uniform extension of the language because we wish to define the meaning of "A + B" for variables which represent Taylor series with real or with interval coefficients.

The support of Pascal-SC for user-defined operators and for overloading is very similar to that provided by ADA. ADA's PACKAGE concept would allow a more secure implementation of data abstractions [10] for real and interval valued Taylor series. The operations on intervals, however, also require support for directed rounding of floating-point results in order to guarantee that the desired answer is contained in the interval computed. The early implementations of ADA do not provide an accuracy of floating-point computations which can compete with Pascal-SC.

Pascal-SC features a highly accurate arithmetic based on a general theory [12] for real and complex numbers, real and complex intervals, and vectors and matrices over these types. Operations on floating-point numbers are rounded to the closest floating-point number to the true result, or upward or downward to the closest neighboring floating-point number under the control of the user. This accuracy meets the proposed IEEE standard for floating-point arithmetic [15]. In addition, scalar products of vectors

$$\text{SCALP(A,B,ROUND)} = \sum_{i=1}^{N} A_i * B_i \tag{2.1}$$

are calculated with the same accuracy (to the closest floating-point number), and with the same options for rounding [24]. A sufficiently long accumulator is used to store intermediate results in the evaluation of the scalar (or inner) product. This capability can also be used to obtain results of the same high accuracy in evaluation of a given arithmetic expression so that 1.0E+99 + 1.0E-99 - 1.0E+99 yields 1.0E-99.

3. TYPES TAYLOR AND ITAYLOR.

We wish to provide the developer of scientific software with a set of tools with which Taylor series methods can be implemented easily for a variety of numerical problems. The ability of the computer to perform formula translation is used. Compilers since the first FORTRAN compiler have produced machine code for the evaluation of an expression such as

(3.1)
```
F := (X*Y + SIN(X) + 4.0) * (3.0 * (Y**2) + 6.0).
```

This is done by analysis of the expression and application of the rules for evaluation of formulas. If the rules for differentiation or recursive generation of Taylor coefficients are applied in the same way, then code for the evaluation of the corresponding quantities results [17]. In this way, fast and inexpensive operations performed by the compiler avoid the overhead involved in invoking symbolic differentiation software. This leads to a more efficient implementation of Taylor series generation all the way from initial coding through program execution.

The normalized Taylor coefficients of a function f(x) expanded at $x = x_0$ are defined by

$$f.TC[K + 1] = f^{(K)}(x_0)t^K/K!, \quad t = x - x_0, \quad K = 0,1,2,\ldots \; . \tag{3.2}$$

Then

$$f_{DIM}(x) = \sum_{K=1}^{DIM} f.TC[K], \tag{3.3}$$

where DIM is the length of the truncated series which is actually stored. This real or interval vector of normalized Taylor coefficients is the basis for the data types TAYLOR and ITAYLOR. For the remainder of this paper, the term "series" is used to refer to the Taylor polynomial given by equation (3.3) or its interval analog.

In what follows, the general rule will be adopted that all variables or expressions of the scalar types INTEGER, REAL, or INTERVAL are treated as constants for the purposes of differentiation.

To form the real data type TAYLOR, the DIM normalized Taylor coefficients in (3.3) are stored as a vector of floating-point numbers. The appropriate declarations in Pascal-SC are:

```
              CONST DIM = n;                         { User supplies n }
              TYPE DIMTYPE  = 1..DIM;
(3.4)              RVECTOR  = ARRAY[DIMTYPE] OF REAL;
                   TAYLOR   = RECORD  LENGTH : DIMTYPE;
                                      T      : REAL;
                                      TC     : RVECTOR  END;
```

These declarations are the same as those given in [19], except for the field named LENGTH. Let F be a variable of type TAYLOR (declared by: VAR F: TAYLOR), then F.LENGTH denotes the actual length of the truncated series ($1 \leq$ F.LENGTH $\leq$ DIM). It may happen that F.LENGTH < DIM if F is being built up recursively, if F has been defined by term-by-term differentiation of another series, or if F has been defined as a quotient of two series both of whose leading terms are zero (see Section 4.2). This field has been added to the record for type TAYLOR given in [19] for internal documentation and so that

i) only series terms actually used need to be processed; and ii) l'Hospital's rule can be applied to certain indeterminant forms 0/0 which may appear.

The normalized Taylor coefficients themselves are stored in the array of real numbers named TC, that is,

$$\text{(3.5)}\qquad \text{F.TC[K]} = F^{(K-1)}(X_0)(X - X_0)^{(K-1)}/K!, \quad K = 1,\ldots,\text{DIM}.$$

The size of the step being used for expansion is $F.T = X - X_0$. Series are generated using a fixed stepsize for which the series might even be divergent. The series for F at a different point Z is readily computed at a cost proportional to DIM:

$$\text{(3.6)}\qquad \text{F.TC[K]} := \text{F.TC[K]}*((Z - X_0)/\text{F.T}))**(K-1); \quad K = 2,\ldots,\text{DIM},$$

while the cost of series generation is usually proportional to DIM^2. The presence of the stepsize in the record also makes it possible to check that an operation is not being performed on two series with different stepsizes.

One of the important problems to which interval analysis has been applied since its beginnings is the problem of controlling the truncation error of Taylor series methods [13]. Hence it is natural to support Taylor series whose normalized coefficients are intervals. The appropriate declarations in Pascal-SC are

```
                CONST DIM = n;                    { User supplies n }
                TYPE DIMTYPE  = 1..DIM;
                     INTERVAL = RECORD INF, SUP : REAL END;
(3.7)                IVECTOR  = ARRAY[DIMTYPE] OF INTERVAL;
                     ITAYLOR  = RECORD  LENGTH : DIMTYPE;
                                        T      : REAL;
                                        TC     : IVECTOR  END;
```

The types ITAYLOR and TAYLOR are the same, except that the normalized coefficients of the former are intervals. The same recurrence relations are used to generate series of each type.

The stepsize T remains real. This corresponds to bounding the range of values of a function f at one real number x. There are some applications for which it is necessary to bound the range of f on an interval, as in (1.13). In this case, one can take T = 1 and form the normalized coefficients by computing the needed powers of (X - x) by interval arithmetic, or else introduce a new data type in which T is of type INTERVAL, and a set of operators corresponding to those given here.

The declarations (3.6) and (3.7) of types TAYLOR and ITAYLOR, respectively, are basic to the discussion of operators in the next section.

4. IMPLEMENTATION OF OPERATORS AND FUNCTIONS FOR TYPES TAYLOR AND ITAYLOR. As indicated above, the ability of a compiler to perform formula translation can also be used to produce machine code for the evaluation of the normalized Taylor coefficients [1], [3], [11], [13], [14], [16], [17], [19]. If the value of function f is obtained by a composition

$$\text{(4.1)}\qquad f = f_1 \circ f_2 \circ \ldots \circ f_m$$

of a finite number of elementary functions, then derivatives of f can be computed by the chain rule from the derivatives of $f_1,\ldots f_m$. This is a tedious and error-prone calculation to do by hand, but the computer does it not only rapidly, but also accurately.

Recurrence relations for calculating the normalized Taylor coefficients for the basic arithmetic operations and for the elementary functions are well known (see [17], for example). Hence machine code can be generated to expand the Taylor series for f at any point $x = x_0$ at which f is analytic. These recurrence relations are both more efficient and more accurate than numerical differentiation [18]. Recursive generation of the series may be mildly unstable [6], but the interval-valued Taylor series introduced in Section 3 can give guaranteed bounds for the effect of any such instability. In Section 6, we show that any instability in the series generation has no significant effect on the series sum.

Rall [19] outlines an approach to abstract data types for real and interval-valued Taylor series. Our implementation generally follows that outline. This paper discusses extensions and some of the implementation details. The operators and functions implemented are listed in Appendix C. Source code in Pascal-SC is given in the report [7]. First of all, in order for expressions to be evaluated correctly when they include variables of type TAYLOR or ITAYLOR, the arithmetic operations and the standard functions must be defined in a manner which incorporates the appropriate recurrence relations for the generation of the normalized Taylor coefficients. Our implementation in Pascal-SC attempts to follow the principles of uniformity, compactness, locality, and linearity for a good programming language design [23]. Next we attempt to justify significant departures from two of these principles.

The principle of uniformity in programming language design says that the same things should be done the same way whenever they occur. Thus "A + B" means "add", regardless of the types of the variables A and B. The other arithmetic operators enjoy the same uniformity, but the standard functions do not. For example, exp(x) is EXP(X) if X is REAL, IEXP(X) if X is INTERVAL, TEXP(X) if X is TAYLOR, and ITEXP(X) if X is ITAYLOR. EXP and IEXP are built-in functions which were designed to suggest the type of their operand and result as an aid to reading the code. That is especially useful since Pascal tends to violate the principle of locality by placing the declaration of a variable far from its use. We chose to maintain uniformity of our extensions with the built-in functions. It is important to be able to determine the type of a variable, and it would be quite non-uniform if IEXP were the only function in this family which requires a prefix.

The principle of locality suggests that all relevant parts of the program are found in the same place. We attempt to follow this principle in each of our program units, but the use of the global constant DIM and the global types RVECTOR, IVECTOR, TAYLOR, and ITAYLOR is a violation. The use of such global types needed in the headings of the operators and functions is very difficult to avoid. Their use has the advantage that all of the information about the length of the series to be used is located in only one place, CONST DIM = n, so it is easy to change.

In roughly their order of importance, the goals of this implementation are:

- Consistent set of software tools.
- Correct answer whenever possible.
- Useful error messages when no correct answer is possible.
- Readable code for future adaptations.
- Efficient execution.
- Compact code.

For example, this implies that although efficient, compact code is sought, efficiency and compactness are sometimes sacrificed for higher goals. In particular, it is important that other programmers be able to read the code, perhaps in order to improve its efficiency.

Binary operations with one operand of type TAYLOR may appear with the other operand of type INTEGER, REAL, or TAYLOR; and the two operands may appear in either order. Similarly, binary operations with one operand of type ITAYLOR can have a second operand of type INTEGER, INTERVAL, or ITAYLOR. The operators built into Pascal-SC do not support the mixing of REAL and INTERVAL operands because real numbers are viewed as being potentially inexact [24]. Our extensions of the arithmetic operators to interval valued Taylor series maintain uniformity with this convention. This is recognized, but not explicitly stated in [19]. If a programmer is certain that a real number X is exact so that it may safely be mixed with an interval, INTPT (X) converts X into the interval [X, X].

The library of subroutines to support computations with types TAYLOR and ITAYLOR includes operators (+,-,*,/,**), special power functions (sqr, sqrt, exp), standard functions (sin, cos, ln, arctan), and additional functions (tan and the Runge function $f(x) = 1/(1 + x^2)$, to which the user can add more functions and procedures as desired. The analytic operations of term-by-term differentiation of real and interval series, as well as term-by-term differentiation of interval series are also provided by means of functions for the given purpose. There is also a set of utility functions and procedures to perform frequently needed tasks, such as reading and writing real and interval series, taking the midpoints of the coefficients of an interval series to obtain a real series, and so on.

The following abbreviations are used in the code to make it as easy as possible to locate a desired operation with any text editor:

K	INTEGER
R	REAL
I	INTERVAL
T	TAYLOR
IT	ITAYLOR.

Using these abbreviations to distinguish between instances of overloading, the operators which are needed to support variables of type TAYLOR and ITAYLOR are:

```
Addition  (Section 4.1):
          + T,  K + T,   T + K,  R + T,   T + R,   T + T
          + IT, K + IT, IT + K,  I + IT, IT + I,  IT + IT
```

```
Subtraction  (Section 4.1):
     -  T,  K -  T,   T - K,  R -  T,   T - R,   T -  T
     - IT,  K - IT,  IT - K,  I - IT,  IT - I,  IT - IT

Multiplication  (Section 4.2):
            K *  T,   T * K,  R *  T,   T * R,   T *  T
            K * IT,  IT * K,  I * IT,  IT * I,  IT * IT

Division  (Section 4.2):
            K /  T,   T / K,  R /  T,   T / R,   T /  T
            K / IT,  IT / K,  I / IT,  IT / I,  IT / IT

Power  (Section 4.3):
     K ** K,    R ** K,  K ** R,    R ** R
                I ** K,  K ** I,    I ** I
     K **  T,   T ** K,  R **  T,   T ** R,   T **  T
     K ** IT,  IT ** K,  I ** IT,  IT ** I,  IT ** IT
```

Implementation details of each operator are discussed in the Sections shown. Pascal-SC provides no power operator, so ** must be defined for the scalar types before it can be extended to types TAYLOR and ITAYLOR. The discussion of ** is postponed to follow the introduction in Section 4.3.1 of special cases of exponentiation: sqr, sqrt, and exp.

The priorities of the operators given in this Section are:

Highest: Unary addition and subtraction, functions;
Multiplication, division, and powers: *, /, **

Lowest: Binary addition and subtraction: +, -

In particular, note that the priority of ** relative to * and / is different than in FORTRAN.

For types TAYLOR and ITAYLOR, implementation has been provided for the standard functions which are supported in Pascal-SC for types INTEGER, REAL, and INTERVAL. They are:

```
Special powers  (Section 4.3.1):
                 TSQR ( T),   TSQRT ( T),   TEXP ( T)
                ITSQR (IT),  ITSQRT (IT),  ITEXP (IT)

Standard functions  (Section 4.4):
            TSIN ( T),   TCOS ( T),   TLN ( T),   TARCTAN ( T)
           ITSIN (IT),  ITCOS (IT),  ITLN (IT),  ITARCTAN (IT)

Additional functions  (Section 4.5):
                      TRUNGE ( T),   TTAN( T)
                     ITRUNGE (IT),  ITTAN(IT)

Differentiation and integration  (Section 4.6):
                      TDIFF( T),   TINTGRL( T)
                     ITDIFF(IT),  ITINTGRL(IT)
```

Miscellaneous utilities (Section 4.7):

```
 VRNULL,    T_IDENT_ZERO( T),    T_IDENT_CONSTANT( T),    ITMIDPT(IT),
IVRNULL,  IT_IDENT_ZERO(IT),  IT_IDENT_CONSTANT(IT),    WRITE_SERIES( T),
READ_INTERVAL SERIES(IT),     WRITE_INTERVAL( I),   WRITE_INTERVAL_SERIES(IT)
```

A brief description of the method for introduction of user-defined functions will be given in Section 4.5. Some implementation details of the operators and functions will now be discussed. The recurrence relations are taken from [17]. In following the conventions of Pascal-SC, minor differences from the indices found there are due to our starting the series indices at 1 instead of starting at 0. In each Section, operations involving the scalar types are discussed before turning to types TAYLOR and ITAYLOR. The Pascal-SC source code for the operators, functions, and utilities listed in Appendix C are given in [7].

4.1. Addition and subtraction.

The ten addition and ten subtraction operators are quite straightforward.

Addition:

```
+  T,   K +  T,    T + K,   R +  T,    T + R,    T +  T
+ IT,   K + IT,   IT + K,   I + IT,   IT + I,   IT + IT
```

Subtraction:

```
-  T,   K -  T,    T - K,   R -  T,    T - R,    T -  T
- IT,   K - IT,   IT - K,   I - IT,   IT - I,   IT - IT
```

Addition and subtraction of a constant alters only the value of a variable, not the values of any of its derivatives. Interval constants only require that the appropriate built-in interval operator be used. Otherwise, addition or subtraction of series is done term-by-term.
If U := F $\pm$ G, then

(4.1.1) $$U.TC[K] := F.TC[K] \pm G.TC[K], \quad K = 1,\ldots,DIM.$$

4.2. Multiplication and Division.

Multiplication:

```
K *  T,    T * K,   R *  T,    T * R,    T *  T
K * IT,   IT * K,   I * IT,   IT * I,   IT * IT
```

Multiplication and division of two Taylor series is done by the well-known Leibniz rule for the Taylor coefficients of a product [17].
If U := F * G, then

(4.2.1) $$U.TC[K] = \sum_{I=1}^{K} F.TC[I]*G.TC[K-I+1], \quad K = 1,\ldots,DIM.$$

The scalar product of two vectors is evaluated in Pascal-SC by the standard function SCALP to the closest floating point number. Fast series multiplication techniques were not used here because

- o In many applications of *, the series for U is being generated recursively. That is, the variables F or G involve U itself.
- o The accuracy of SCALP would not be available.
- o The speed of SCALP, especially when some terms are zero, makes these techniques less attractive.

Multiplication or division of a series by a constant is done term-by-term. Division of a constant by a series is done by generation of the series for C/F(x).

Division:

K / T, T / K, R / T, T / R, T / T
K / IT, IT / K, I / IT, IT / I, IT / IT

If U := F / G, then U * G = F, and Leibniz' rule applies:

$$\begin{aligned} &\mathrm{U.TC}[1] = \mathrm{F.TC}[1] \,/\, \mathrm{G.TC}[1], \\ &\text{for } K = 2,\ldots,\mathrm{DIM}, \\ &\mathrm{U.TC}[K] = \Big(\sum_{I=1}^{K-1} \mathrm{U.TC}[I]*\mathrm{G.TC}[K-I+1] \Big) \,/\, \mathrm{G.TC}[1], \end{aligned} \tag{4.2.2}$$

If $G(x_0)$ = G.TC[1] = 0, then we attempt return the correct answer whenever possible. If $F(x_0)$ = F.TC[1] is also 0, then we can apply l'Hospital's rule because the series for both F and G are known. U.TC[1] = $F'(x_0)/G'(x_0)$ = F.TC[2]/G.TC[2], if this quotient exists, but U ≠ F'/G' as functions.
If U := F/G, and $F(x_0) = G(x_0) = 0$, then let

$$\mathrm{V.TC}[K] := \mathrm{F.TC}[K+1]; \quad \mathrm{W.TC}[K] := \mathrm{G.TC}[K+1]; \quad K = 1,\ldots,\mathrm{DIM}-1. \tag{4.2.3}$$

Then,

$$U := V \,/\, W. \tag{4.2.4}$$

Thus, l'Hospital's rule is implemented as a recursive call to the division operator with operands whose series length has been reduced by one. This approach would not be possible in a language which does not support recursion. Further, cases in which the series for both f and g have several leading zeros are handled automatically by the language.

L'Hospital's rule is applied in a similar manner when a constant quotient or divisor is equal to zero.

4.3. Power Operators. The power operator ** defined by F ** G = F^G is not standard in Pascal or Pascal-SC, but can be implemented in the latter for data types for which it is meaningful by the use of the operator concept. Coding of ** is simplified by the introduction of a set of basic power functions. These are implemented separately

- for uniformity with Pascal-SC which provides these functions for standard data types,
- to provide tighter bounds for interval operands, and
- for efficiency.

4.3.1. Special Power Functions This set of functions consists of the square, square root, and natural exponential function of variables of types TAYLOR and ITAYLOR:

TSQR (T), TSQRT (T), TEXP (T)
ITSQR (IT), ITSQRT (IT), ITEXP (IT)

These functions are called by the operator ** when appropriate. For example, if X is of type ITAYLOR, then both X ** 2 and X ** INTPT(2.0) are actually performed by a call to ITSQR (X). The use of this function rather than X * X is important in interval computations, since, for example, $[-1,+1]^2 = [0,1]$ while [-1,+1] * [-1,+1] = [-1,+1]. Further, the squaring functions TSQR and ITSQR are twice as fast as the multiplication Y * Y for variables of the corresponding types.

The recurrence relations to generate the series terms for these functions can be derived easily using Leibniz' rule. The squares of real and interval Taylor series are computed as follows.
If U := SQR(F), then Leibniz' rule for a product can be shortened to:

```
            For K = 1, ..., DIM,

(4.3.1)     $U.TC[K] = \sum_{I = 1}^{K \text{ DIV } 2} F.TC[I]*F.TC[K-I+1];$

            if K is odd, then U.TC[K] = U.TC[K] + SQR (F.TC[(K+1)/2]).
```

The inner product contains only TRUNC(K/2) terms. If F is of type ITAYLOR and includes negative numbers, then ITSQR (F) provides tighter bounds than does F * F. The SQR functions are named TSQR and ITSQR to indicate the type of operand accepted and value returned.

A similar function was written for CUBE. Its summations had length TRUNC(K/3) but they were nested to yield a cost proportional to DIM^3. CUBE is not included in the library because F * SQR (F) is faster.

The functions in the next set calculate square roots of real and interval Taylor variables.
If U := SQRT (F), then U * U = F. The algorithm runs as follows:

```
               U.TC[1] := SQRT (F.TC[1]);

               U.TC[2] := F.TC[2] / (2 * U.TC[1]);

               for K = 3, ..., DIM,

(4.3.2)            $PROD := \sum_{I = 2}^{K \text{ DIV } 2} U.TC[I]*U.TC[K-I+1];$

               if K is odd, then PROD := PROD + SQR (U.TC[(K+1)/2]);

               U.TC[K] := (F.TC[K] - PROD) / (2 * U.TC[1]).
```

If $F(x_0) = 0$, and F is not a constant series, then SQRT (F) cannot be computed unless F is the constant 0, because $F'(x_0)$ is not defined. The SQRT functions are named TSQRT and ITSQRT to indicate the type of operand accepted and value returned.

The natural exponential functions (base = e) are now defined for types TAYLOR and ITAYLOR.
If U := EXP (F), then U' = U * F'. This gives the algorithm:

```
            U.TC[1] := EXP (F.TC[1]);

            for K = 2,...,DIM,

(4.3.3)
            U.TC[K] := ( $\sum_{I=1}^{K-1}$ U.TC[I]*F.TC[K-I+1]*(K-I) ) / (K-1);
```

Although this formula would appear slightly simpler with the change of index J = K - I, it was implemented in this way so that the U.TC terms remain stationary in the inner product as K increases. Thus, only the vector F.TC needs to be "reversed". The EXP functions are named TEXP and ITEXP to indicate the type of operand accepted and value returned.

4.3.2. The Operator **. The family of power operators ** seems to be the most difficult to implement as suggested by [5]. None of the operators are especially difficult, but there are many minor details to be considered. The implementation of ** for types TAYLOR and ITAYLOR is based on the standard power functions above, and the power operators ** for the scalar types INTEGER (K), REAL (R), and INTERVAL (I).

```
    Scalar Powers:
            K ** K,    R ** K,  K ** R,    R ** R
                       I ** K,  K ** I,    I ** I
```

Integer powers are implemented using repeated squaring. Real and interval powers which fit no special case are computed by F ** G = EXP (G * LN(F)). We have not attempted optimal implementations of the scalar power operators because it is hoped that they will be provided as standard operators in a later release of Pascal-SC, an approach that is especially attractive for interval operands because the interpreter hides information from programmers which can be used for correctly directed roundings.

```
    Real and Interval Taylor Powers:
            K **  T,   T ** K,  R **  T,   T ** R,   T **  T
            K ** IT,  IT ** K,  I ** IT,  IT ** I,  IT ** IT
```

The power operators for a constant to a variable power follow the pattern of TEXP or ITEXP, as appropriate. A series which represents a constant (only its first term is non-zero) is handled as a special case for accuracy (especially for interval series) and for efficiency.

The operators T ** K and IT ** K take care to return the correct answer whenever that is possible and to produce an appropriate error message when it is not possible. The resolution of various cases is shown in Table 4.1.

Consider a series whose first term is zero, but which has other terms which are non-zero. Raising such a series to a negative power is undefined because it is equivalent to dividing by zero, but raising such a series to the power 0 defines a function which is identically equal to 1, except for a removable singularity at $x = x_0$. Hence it is appropriate to give 1 as the answer. Raising the series to a positive integer power is implemented by repeated squaring because the recurrence relation which is most often used [21] requires division by BASE.TC[1], which is zero.

Exponent:	0	1	2	> 2	< 0
Base.TC = 0	Undef.	= 0	= 0	= 0	Undef.
Base.TC[1] = 0	1	= BASE	SQR (BASE)	By mult.	Undef.
Base.TC[1] <> 0	1	= BASE	SQR (BASE)	By recurrence	

Table 4.1. Resolution of Cases for **.

The special cases of an exponent equal to 1 or 2 are singled out for individual treatment in order to achieve the maximum possible accuracy (especially when the base is an interval series) and for efficiency.

Except in the special cases shown in the table, if U = F ** E, where E is of type INTEGER, then F * U' = E * U * F'. This gives the algorithm

```
              U.TC := 0;

              U.TC[1] := F.TC[1] ** E;

(4.3.4)       For K := 2 to DIM,

   $U.TC[K] := \left( \sum_{I=1}^{K-1} (E*(K-I) - I + 1) * U.TC[I] * F.TC[K-I+1] \right) / ((K-1)*F.TC[I]).$
```

The integer exponent appears in the recurrence only as a multiplier. Hence the speed of this algorithm is nearly independent of the size of the exponent. That is why this algorithm is preferred to repeated squaring.

The operators T ** R and IT ** I are similar to T ** K and IT ** K, respectively, except that the additional special cases of an exponent equal to 1/2 or to an integer are handled.

The operators T ** T and IT ** IT are included primarily for completeness; the authors have never seen a differential equation containing a variable to a variable power, for example. Perhaps any such problems which arise are at once simplified by logarithmic differentiation. With the tools described here, it may be advantagious to attack the problem in its original form.

Within the operators T ** T and IT ** IT, the cases in which either the base or the exponent series represent a constant function are treated as special for reasons of accuracy and efficiency. Otherwise, F ** G = EXP (G * LN (F)), using TEXP and TLN or ITEXP and ITLN, as appropriate.

4.4. Standard Functions. There are many useful library functions which could be provided. We have chosen to implement the functions which are built into Pascal-SC for the standard scalar data types, and a few others. Additional functions can be added as they are needed by following the models provided by this paper. In addition to the standard power functions of Section 4.3.1 (which include EXP and IEXP), other standard functions implemented for types TAYLOR and ITAYLOR are:

TSIN (T), TCOS (T), TLN (T), TARCTAN (T)
ITSIN (IT), ITCOS (IT), ITLN (IT), ITARCTAN (IT)

If U = sin (F) and V = cos (F), then U' = V * F' and V' = - U * F'.

$$\text{U.TC[1]} := \text{SIN (F.TC[1]);} \quad \text{V.TC[1]} := \text{COS (F.TC[1]);}$$

$$\text{for K} := 2,\ldots,\text{DIM},$$

$$\text{U.TC[K]} := \Big(\sum_{I=2}^{K} \text{V.TC[I]} * \text{F.TC[K-I+1]} * (\text{K}-\text{I}) \Big) / (\text{K}-1); \tag{4.4.1}$$

$$\text{V.TC[K]} := - \Big(\sum_{I=2}^{K} \text{U.TC[I]} * \text{F.TC[K-I+1]} * (\text{K}-\text{I}) \Big) / (\text{K}-1).$$

The SIN and COS functions are named TSIN and TCOS or ITSIN and ITCOS to indicate the type of operand which they accept and value they return. Since the series for SIN and COS are always computed together, the library also contains procedures T_SIN_COS and IT_SIN_COS which return both the SIN and COS of variables of type TAYLOR and ITAYLOR, respectively, in the same call.
If U := ln (F), then U' * F = F'.

$$\text{U.TC[1]} := \text{LN (F.TC[1]);}$$

$$\text{for K} := 2,\ldots,\text{DIM}, \tag{4.4.2}$$

$$\text{U.TC[K]} := \Big(\text{F.TC[K]} - \Big(\sum_{I=2}^{K-1} \text{U.TC[I]} * \text{F.TC[K-I+1]} * (\text{I}-1) \Big) / (\text{K}-1)\Big) / \text{F.TC[1]}.$$

There is a misprint in this recurrence relation in ([17], p. 42), but its implementation is straightforward.
If U := arctan (F) and V := 1 / (1 + F^2), then U' = V * F'.

$$\text{V} := 1 / (1 + \text{SQR (F)});$$

$$\text{U.TC[1]} := \text{ARCTAN (F.TC[1]);}$$

$$\text{for K} := 2,\ldots,\text{DIM}, \tag{4.4.3}$$

$$\text{U.TC[K]} := \Big(\sum_{I=2}^{K} \text{V.TC[I]} * \text{F.TC[K-I+1]} * (\text{K}-\text{I}) \Big) / (\text{K}-1).$$

Since the series for the Runge function V(F) ([8], p. 78) is required to compute the series for arctan(F), functions TRUNGE and ITRUNGE are included in the library along with the functions TARCTAN and ITARCTAN.

4.5. User Defined Functions. If a programmer requires an operation or a function which is not included in this report, the requirement can be met either by a composition of operators and functions which are already provided, or by a careful derivation of the necessary recurrence relations following the models in this report. For example, the tangent functions TTAN(T) and ITTAN(IT) are implemented essentially by

(4.5.1)

TTAN(T) := TSIN(T) / TCOS(T),

ITTAN(IT) := ITSIN(IT) / ITCOS(IT),

respectively, in the set of additional functions provided in the library. The tangent functions can also be implemented directly by recurrence relations, using the fact that y = tan(x) satisfies the differential equation

$$y' = 1 + y^2, \quad y(x_0) = \tan(x_0), \tag{4.5.2}$$

[13], [14]. Thus, for U := TTAN(T), for example,

(4.5.3) U.TC[1] := TAN(T.TC[1]), U.TC[2] := (1 + SQR(T.TC[1])) * T.T.

The succeeding coefficients can be obtained in a simple way from the recurrence relation (4.3.1) for TSQR(T), and ITTAN(IT) is computed analogously.

The Runge function $f(x) = 1 / (1 + x^2)$, which is an auxiliary function for the series expansion of the arctangent, is implemented in the library by

U := TSQR(IT);

(4.5.4) U.TC[1] := 1 + U.TC[1];

TRUNGE := 1 / U.

ITRUNGE is computed similarly.

4.6. Differentiation and Integration. Functions which return the results of term-by-term differentiation and integration of TAYLOR and ITAYLOR series are also provided. For series with 1 < LENGTH $\leq$ DIM, differentiation decreases the length of the series by one:

(4.6.1) U.TC[K] := T.TC[K + 1] * RATIO / (K + 1), K = 1,...,T.LENGTH - 1,

where RATIO = 1 / T.T if U = TDIFF (T), and RATIO = 1 / INTPT (T.T) if U = ITDIFF (T). Integration results in a series with its first coefficient set to 0 and its length increased by one:

(4.6.2) U.TC[K] := T.TC[K - 1] * RATIO / (K - 1), K = 2,...,T.LENGTH,

with U.TC[1] = 0 and RATIO = T.T for U = TINTGRL(T), while U.TC[1] = INTPT(0) and RATIO = INTPT(T.T) for U = ITINTGRL(T). The result of integration of a series of length DIM will be truncated to length DIM.

4.7. Miscellaneous Utilities. Some useful functions and procedures are provided for convenience. These are the transfer function TMIDPT (IT), the special functions VRNULL, IVRNULL, the comparison functions T_IDENT_ZERO, T_IDENT_CONSTANT, IT_IDENT_ZERO, IT_IDENT_CONSTANT, and the input/output procedures WRITE_INTERVAL (I), WRITE_SERIES (T), READ_INTERVAL_SERIES (IT), WRITE_INTERVAL_SERIES (IT). The purposes of most of these utilities are indicated by their names.

The transfer function TMIDPT (IT) forms a TAYLOR series from a series of type ITAYLOR. The coefficients of the result series are the midpoints of the corresponding coefficients of the interval series.

The parameterless functions VRNULL, IVRNULL yield zero real and interval vectors, respectively, of length DIM. They are standard in Pascal-SC.

The comparison functions yield the BOOLEAN value TRUE if their argument satisfies the stated condition (the series is identically equal to zero or a constant), otherwise, FALSE.

The input/output procedures are also self-explanatory. The procedure WRITE INTERVAL is included, since the standard Pascal-SC procedure IWRITE only prints the digits of the lower and upper endpoints of intervals which agree up to the last [24]. WRITE_INTERVAL, however, prints all digits of each endpoint.

5. THE INITIAL-VALUE PROBLEM FOR ORDINARY DIFFERENTIAL EQUATIONS. Taylor series methods for the numerical solution of initial-value problems for systems of ordinary differential equations have been studied by many authors (see [6] or [14] for summaries), and have been used for applications such as dynamics and parameter identification. Each component of the solution of

$$(5.1) \qquad y_i' = f_i(x,y), \quad y_i(x_0) = y_{i0}, \quad i = 1,\ldots,m,$$

is expressed as a Taylor series expanded at $x = x_0$ using recurrence relations derived from the differential equation. Various error control strategies have been employed. The strategy of analyzing the radius of convergence of each component series has the desirable side effect of producing such analytic information as the location and orders of the singularities in the solution. Once the radius of convergence is known, a stepsize can be chosen which is as large as possible subject to error control and stability constraints. Then each component of the solution is extended by analytic continuation and the process is repeated at the next integration step. This algorithm is discussed in greater detail in [6].

A program RDEQ_SOLV for solving equation (5.1) is given as Appendix A of this report. The program is written for the case m = 1, but can be modified easily to handle systems of several equations. The variables Y and YPRIME are declared to be of type TAYLOR, and the equation is written in a natural way. To solve a different equation, it is only necessary

- o to change the line in RDEQ_SOLV which contains the differential equation;
- o to copy from the library into the source program any operators or functions required by the new differential equation.

Because the differential equation is written using the types and operators discussed in the preceding Sections, the needed recurrence relations are implemented by the Pascal-SC compiler and need not be derived explicitly by the user.

The program prints the series terms, extends the solution by analytic continuation to compute the initial condition at the next step, and repeats the process. The program RDEQ_SOLV in Appendix A is intentionally simple to illustrate the use of the Taylor operators and to explore the stability of the series generation. It would require an error control mechanism in order to be of practical value for the solution of initial value problems. Either scalar [6] or interval [13] error control techniques can be used.

The program IDEQ_SOLVE listed as Appendix B of this report computes interval-valued approximate solutions to equation (5.1) for the case m = 1, but can be modified for systems of several equations. It differs from the program RDEQ_SOLV only in that

i) the variables Y and YPRIME are of type ITAYLOR instead of type TAYLOR, and

ii) additional code has been added to monitor the relative error introduced by instability in the series generation process.

These two programs are designed to serve as examples of one way in which the tools of this report can be used. They are simple, menu-driven programs which allow direct user intervention at each integration step. By observation of the outcome of each step, the user can experiment with error control strategies.

The bounds computed by IDEQ_SOLV are for the interval-valued Taylor polynomial (3.3). They are not global error bounds for the solution of the differential equation. Global error bounds are readily computable using interval remainder terms (see [13]), but, for simplicity, the programs given here contain no error bounding or control strategy.

6. AN APPLICATION: STABILITY OF SERIES GENERATION.

In this section, we present an example which uses the Taylor and interval Taylor operators. This example was chosen because it illustrates the uses of these operators and because it addresses the issue of stability in the generation of the series. The latter issue is central in showing that Taylor series methods are reliable for practical computations.

A numerical computation is said to be unstable if its relative error grows without bound as the computation proceeds. It is possible that the recurrence relations being used might be unstable, although instability has never been observed in practice. This example uses the Taylor and interval Taylor operators to explore the stability of the recurrence relations in one application. In this example, there is instability in the generation of the terms of the series, but that does not seriously affect the accuracy of the series summation. The stability of the recurrence relations in other applications can be handled similarly.

Consider the initial value problem

$$y' = y^2, \quad y(0) = 1, \tag{6.1}$$

whose solution is $y(x) = 1 / (1 - x)$. A program (RDEQ_SOLV) for solving equation (6.1) using the Taylor function TSQR is given as Appendix A of this report. The effect of program RDEQ_SOLV is to generate the normalized Taylor coefficients (4.3.1) of the solution recursively. This recurrence is accomplished automatically by the Taylor function TSQR in the statement YPRIME := TSQR (Y). In this case, the same solution is obtained if Y ** 2 or Y * Y is used instead of TSQR (Y); however, the use of Y ** 2 requires the compilation of much more code, while Y * Y is not as fast as TSQR (Y).

We wish to explore the stability of the recurrence relation (4.3.1). This issue is separate from the issue of the stability of Taylor series methods for solving initial value problems. If an infinite Taylor series were used, the method would be A-stable, but when a truncated series is used, the region of stability is bounded. Stetter [22] showed that the region of stability for truncated Taylor series methods is the same as that for related Runge-Kutta methods. The real interval of stability is relatively large here because long series are used. For example, the real intervals of stability are [-8.85, 0] and [-16.29, 0], respectively, if DIM = 20 and 40 terms of the series are used.

We will outline the theoretical analysis of the stability of recurrence (4.3.1). A more complete discussion appears in [6]. Let U(K) denote the actual and Y(K) denote the computed normalized Taylor coefficients. Let Y(1) = U(1) $(1 + \varepsilon) = (1 + \varepsilon)$ from (6.1). Then U(K) = h^{K-1}, and

$$Y(K) = U(K)\ (1 + \varepsilon)^K, \tag{6.2}$$

so the series generation is unstable. However, the summation of the series is unaffected by this instability since

$$\begin{aligned} y(x_1) &= \sum_{K=1}^{DIM} Y(K) = (1 + \varepsilon) \sum_{K=1}^{DIM} (h(1 + \varepsilon))^{K-1} . \\ &= (1 + \varepsilon)\ y(h(1 + \varepsilon)) + O(h^{DIM}) \\ &= (1 + \varepsilon)(y(h) + y'(\xi)h) + O(h^{DIM}), \quad h < \xi < h(1 + \varepsilon). \end{aligned} \tag{6.3}$$

That is, the instability in the series generation is equivalent to a small error in the point at which the series is evaluated. This is because $\Sigma_K Y(K)$ is a convergent series, so the terms for which instability causes the relative error to be largest are themselves very small.

This suggests using interval arithmetic to keep track of the potential growth in the series. The program IDEQ_SOLV listed as Appendix B of this report does so.

By declaring Y and YPRIME to be of type ITAYLOR, the statement YPRIME := ITSQR(Y) invokes the function ITSQR for interval valued series to generate interval normalized Taylor coefficients according the recurrence relation (4.3.1). The lengths of successive coefficients measure the stability of the recurrence. Table 6.1 shows the interval valued series solution of equation (6.1) for DIM = 15, $y(x_0) = y(0) = [0.99\ ,\ 1.01]$ ($\varepsilon = \pm\ 0.01$), and h = 0.5.

```
READ INTERVAL INITIAL CONDITIONS X0, Y(X0):


INITIAL CONDITIONS AT X0 = [ 0.00000E+00,  0.00000E+00],
                      Y0 = [ 9.90000E-01,  1.01000E+00].
ENTER STEPSIZE X - X0:  0.5
Computing series terms ...


        Step     Left          Right         Computed       Theoretical
               Endpoint       Endpoint      Instability     Instability

          1  [ 9.90000E-01,  1.01000E+00]    1.010E+00       1.010E+00
          2  [ 4.90050E-01,  5.10050E-01]    1.020E+00       1.020E+00
          3  [ 2.42575E-01,  2.57575E-01]    1.030E+00       1.030E+00
          4  [ 1.20075E-01,  1.30076E-01]    1.040E+00       1.041E+00
          5  [ 5.94369E-02,  6.56881E-02]    1.050E+00       1.051E+00
          6  [ 2.94213E-02,  3.31725E-02]    1.060E+00       1.062E+00
          7  [ 1.45635E-02,  1.67521E-02]    1.070E+00       1.072E+00
          8  [ 7.20894E-03,  8.45982E-03]    1.080E+00       1.083E+00
          9  [ 3.56843E-03,  4.27221E-03]    1.090E+00       1.094E+00
         10  [ 1.76637E-03,  2.15747E-03]    1.100E+00       1.105E+00
         11  [ 8.74354E-04,  1.08952E-03]    1.110E+00       1.116E+00
         12  [ 4.32805E-04,  5.50208E-04]    1.119E+00       1.127E+00
         13  [ 2.14239E-04,  2.77855E-04]    1.129E+00       1.138E+00
         14  [ 1.06048E-04,  1.40317E-04]    1.139E+00       1.149E+00
         15  [ 5.24938E-05,  7.08599E-05]    1.149E+00       1.161E+00

THE VALUE AT X = [ 5.00000E-01,  5.00000E-01]
          IS Y = [ 1.96034E+00,  2.04033E+00].
```

Table 6.1. Interval bounds for instability.

The computed instability is equal to

$$E_{\text{Computed}} = \frac{\text{length (Y.TC[K])}}{\text{midpoint (Y.TC[K])}}, \tag{6.4}$$

a measure of the relative error in Y.TC(K) which appears to grow as K increases. The theoretical instability is equal to

$$E_{\text{Theoretical}} = (1 + \varepsilon)^{K}. \tag{6.5}$$

Table 6.1 shows that these two values are very close, and that the theoretical bound is slightly larger than the actual bound, as it should be. The interval estimate for y(0.5) agrees well with the interval [y(0.46), y(0.54)] = [1.8518, 2.1739].

7. <u>IMPLEMENTATION DETAILS</u>. The software described in this report was created and tested using the Pascal-SC compiler developed at the University of Karlsruhe for the Zilog MCZ-1 microcomputer with the RIO 2.06 operating system. No other claims of correctness or usability are made.

REFERENCES

1. D. Barton, I. M. Willers, and R. V. M. Zahar. Taylor series methods for ordinary differential equations - An evaluation. In Mathematical Software, John Rice (Ed.). Academic Press, New York, 1971, 369-390.

2. G. Bohlender, K. Gruner, E. Kaucher, R. Klatte, W. Kramer, U. W. Kulisch, S. M. Rump, Ch. Ullrich, J. Wolff von Gudenberg, and W. L. Miranker. Pascal-SC: A Pascal for contemporary scientific computation. Research Report RC 9009, IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y., 1981.

3. Y. F. Chang. Automatic solution of differential equations. In Constructive and Computational Methods for Differential and Integral Equations, D. L. Colton and R. P. Gilbert (Eds.). Lecture Notes in Mathematics, Vol 430, Springer-Verlag, Berlin-Heidelberg-New York, 1974, 61-94.

4. Y. F. Chang, M. Tabor, J. Weiss, and G. F. Corliss. On the structure of the Henon Heiles system. Phys. Lett. A 85A (1981), 211-213.

5. W. J. Cody and W. Waite. Software Manual for the Elementary Functions. Prentice-Hall, Englewood Cliffs, N. J., 1980.

6. George Corliss and Y. F. Chang. Solving ordinary differential equations using Taylor series. ACM Trans. Math. Soft. 8 (1982), 114-144.

7. George Corliss and L. B. Rall. Automatic Generation of Taylor Series in Pascal-SC: Basic Operations and Applications to Ordinary Differential Equations. MRC Technical Summary Report No. 2497, Mathematics Research Center, University of Wisconsin-Madison, 1983.

8. Philip J. Davis. Interpolation and Approximation. Blaisdell, New York, 1963.

9. K. Jensen and N. Wirth. Pascal User Manual and Report, 2nd Ed. Springer-Verlag, Berlin-Heidelberg-New York, 1974.

10. Carlo Ghezzi and Mehdi Jazayeri. Programming Language Concepts. Wiley, New York, 1982.

11. G. Kedem. Automatic differentiation of computer programs. ACM Trans. Math. Soft. 6 (1980), 150-165.

12. U. W. Kulisch and W. L Miranker. Computer Arithmetic in Theory and Practice. Academic Press, New York, 1981.

13. R. E. Moore. Interval Analysis. Prentice-Hall, Englewood Cliffs, N. J., 1966.

14. R. E. Moore. Methods and Applications of Interval Analysis. SIAM Studies in Applied Mathematics, 2, Philadelphia, 1979.

15. J. F. Palmer. VSLI and the revolution in numeric computation. Proceedings of the 10th IMACS World Congress on System Simulation and Scientific Computation, Vol. 1, pp. 339-341. Montreal, 1982.

16. L. B. Rall. Applications of software for automatic differentiation in numerical computation. Computing, Suppl. 2 (1980), 141-156.

17. L. B. Rall. Automatic Differentiation: Techniques and Applications. Lecture Notes in Computer Science No. 120, Springer-Verlag, Berlin-Heidelberg-New York, 1981.

18. L. B. Rall. Differentiation in Pascal-SC: Type GRADIENT. Technical Summary Report No. 2400, Mathematics Research Center, University of Wisconsin-Madison, 1982.

19. L. B. Rall. Differentiation and generation of Taylor coefficients in Pascal-SC. Technical Summary Report No. 2452, Mathematics Research Center, University of Wisconsin-Madison, 1982.

20. L. B. Rall. Mean-value and Taylor forms in interval analysis. SIAM J. Math. Anal. 14 (1983), no. 2, 223-238.

21. Allen Reiter. Automatic generation of Taylor coefficients (TAYLOR) for the CDC 1604. Technical Summary Report No. 830, Mathematics Research Center, University of Wisconsin-Madison, 1967.

22. H. J. Stetter. Analysis of Discretization Methods for Ordinary Differential Equations. Springer-Verlag, Berlin-Heidelberg-New York, 1973.

23. G. Weinberg. The Psychology of Computer Programming. Van Nostrand Reinhold, New York, 1971.

24. J. Wolff von Gudenberg. Gesamte Arithmetik des PASCAL-SC Rechners: Benutzerhandbuch. Institute for Applied Mathematics, University of Karlsruhe, 1981.

APPENDIX A

```
PROGRAM RDEQ_SOLVE (INPUT, DATA, OUTPUT);

(*  SOLVE A FIRST ORDER DIFFERENTIAL EQUATION:  Y' = SQR (Y)  *)

CONST DIM = 30;
TYPE DIMTYPE  = 1..DIM;
     RVECTOR  = ARRAY[DIMTYPE] OF REAL;
     TAYLOR   = RECORD  LENGTH : DIMTYPE;
                        T      : REAL;
                        TC     : RVECTOR  END;
     CHOICE   = 1..3;

VAR     FLAG            : CHOICE;
        I, IM1          : DIMTYPE;
        X, Y, YPRIME    : TAYLOR;
        DATA            : TEXT;
FUNCTION VRNULL : RVECTOR;
    VAR I: DIMTYPE; U: RVECTOR;
    BEGIN
    FOR I := 1 TO DIM DO U[I] := 0.0;
    VRNULL := U
    END;  (*  FUNCTION VRNULL  *)
FUNCTION TSQR (T: TAYLOR) : TAYLOR;                              (*  TSQR(T)  *)
    (*  Requires:  VRNULL, SCALP, SQR  *)
    VAR I, J, K, HALF: DIMTYPE;
        X, Y: RVECTOR;
        U   : TAYLOR;
    BEGIN
    X := VRNULL;  Y := VRNULL;
    U.LENGTH := T.LENGTH;
    U.T := T.T;
    U.TC := VRNULL;
    U.TC[1] := SQR (T.TC[1]);
    X[1] := T.TC[1];
    FOR K := 2 TO U.LENGTH DO
        BEGIN
        X[K] := T.TC[K];
        HALF := K DIV 2;
        FOR J := 1 TO HALF DO
            BEGIN
            I := K - J + 1;
            Y[J] := T.TC[I]
            END;  (* FOR J *)
        U.TC[K] := 2.0 * SCALP ( X, Y, 0);
        IF K MOD 2 = 1 THEN
            BEGIN
            HALF := HALF + 1;
            U.TC[K]:= U.TC[K] + SQR (T.TC[HALF])
            END  (* IF *)
        END;  (* FOR K *)
    TSQR := U
    END;  (* FUNCTION TSQR (TAYLOR)  *)
```

```
FUNCTION MENU_CHOICE : CHOICE;
    VAR I: INTEGER;
    BEGIN
    WRITELN;
    WRITELN ('ENTER:  1   -  GIVE NEW INITIAL CONDITIONS');
    WRITELN ('        2   -  CONTINUE EXTENDING THE SOLUTION');
    WRITELN ('        3   -  STOP');
    READ (I);
    IF ((I >= 3) OR (I <= 0)) THEN I := 3;
    MENU_CHOICE := I
    END;  (*  FUNCTION MENU_CHOICE  *)
PROCEDURE PRNT_TAY_COEF (Y: TAYLOR; INDEX: DIMTYPE);
    BEGIN
    WRITELN   ('Y(', INDEX:5, ') = ', Y.TC[INDEX])
    END;  (*  PROCEDURE PRNT_TAY_COEF  *)
FUNCTION SUM (VAR A: RVECTOR; DIM, ROUND: INTEGER) : REAL;
  EXTERNAL 480;

BEGIN  (*MAIN PROGRAM RDEQ_SOLVE*)
(*...........  INITIALIZE *)
FLAG := 2;
X.LENGTH := DIM;
Y.LENGTH := DIM;

RESET (DATA);
WHILE FLAG <= 2 DO              (* LOOP FOR NEW INITIAL CONDITIONS *)
    BEGIN
    FLAG := 2;
X.TC := VRNULL;
Y.TC := VRNULL;

    WRITELN ('READ REAL INITIAL CONDITIONS X0, Y(X0):');
    READ (DATA, X.TC[1]); READ (DATA, Y.TC[1]);
    WRITELN;  WRITELN;
    WRITELN ('INITIAL CONDITIONS AT X0 = ', X.TC[1], ',');
    WRITELN ('                      Y0 = ', Y.TC[1], '.');
    WHILE FLAG = 2 DO           (* LOOP FOR ANALYTIC CONTINUATION *)
        BEGIN
        (*..........  READ STEP SIZE *)
        WRITELN ('ENTER STEPSIZE X - X0:  ');
        READ (X.T);
        Y.T := X.T;
        WRITELN ('Computing series terms ...');
        FOR I := 2 TO DIM DO            (* LOOP FOR SERIES GENERATION *)
            BEGIN

(*      YOUR FIRST ORDER DIFFERENTIAL EQUATION GOES HERE:          *)
    YPRIME := TSQR (Y);

            IM1 := I - 1;
            Y.TC[I] := YPRIME.TC[IM1] * Y.T / IM1;
            END;  (*FOR*)
        (*..........  PRINT TABLE *)
        WRITELN;  WRITELN;
        WRITELN ('THE TAYLOR COEFFICIENTS OF Y ARE:');
        FOR I := 1 TO DIM DO PRNT_TAY_COEF (Y, I);
        (*.......... PERFORM THE ANALYTIC CONTINUATION *)

        Y.TC[1] := SUM (Y.TC, DIM, 0);
        FOR I:= 2 TO DIM DO Y.TC[I] := 0.0;
        X.TC[1] := X.TC[1] + X.T;
        WRITELN;
        WRITELN ('THE VALUE AT X = ', X.TC[1]);
        WRITELN ('          IS Y = ', Y.TC[1], '.');
        FLAG := MENU_CHOICE
        END  (*WHILE*)

    END  (*WHILE*)

END.  (*  MAIN PROGRAM RDEQ_SOLVE  *)
```

APPENDIX B

```
PROGRAM IDEQ_SOLVE (INPUT, DATA, OUTPUT);

(*  SOLVE A FIRST ORDER DIFFERENTIAL EQUATION   *)
(*  Y' = SQR (Y)   *)
(*  SOLUTION IS IN INTERVAL FORM   *)

CONST DIM = 15;

TYPE DIMTYPE  = 1..DIM;
     INTERVAL = RECORD INF, SUP : REAL END;
     IVECTOR  = ARRAY[DIMTYPE] OF INTERVAL;
     ITAYLOR  = RECORD  LENGTH : DIMTYPE;
                        T      : REAL;
                        TC     : IVECTOR  END;
     CHOICE   = 1..3;

VAR     FLAG              : CHOICE;
        I, IM1            : DIMTYPE;
        X, Y, YPRIME      : ITAYLOR;
        DATA              : TEXT;
        EPSILON,
        COMPOUND          : REAL;

       (*  Transfer Functions   *)

FUNCTION INTPT ( RA:REAL ) : INTERVAL;
  EXTERNAL 41;
FUNCTION INTVAL ( RA,RB: REAL ) : INTERVAL;
  EXTERNAL 42;
FUNCTION IINF ( A: INTERVAL) : REAL;
  EXTERNAL 43;
FUNCTION ISUP ( A: INTERVAL) : REAL;
  EXTERNAL 44;

        (* Comparisons *)

OPERATOR <= (A,B: INTERVAL ) RES: BOOLEAN ;
  EXTERNAL 48;
OPERATOR >= (A,B: INTERVAL ) RES: BOOLEAN ;
  EXTERNAL 50;
OPERATOR IN (RA:REAL; B: INTERVAL) RES: BOOLEAN;
  EXTERNAL 47;
OPERATOR IN (KA: INTEGER; B: INTERVAL) RES: BOOLEAN;
  EXTERNAL 46;
OPERATOR >< (A,B: INTERVAL ) RES: BOOLEAN ;
  EXTERNAL 52;

        (* Lattice Operators *)

OPERATOR ++ (A,B: INTERVAL) RES: INTERVAL;
  EXTERNAL 63;
OPERATOR ** (A,B: INTERVAL) RES: INTERVAL;
  EXTERNAL 60;
```

```
        (* Arithmetic Operators *)

OPERATOR + ( A: INTERVAL ) RES: INTERVAL;
  EXTERNAL 67;
OPERATOR - ( A: INTERVAL ) RES: INTERVAL;
  EXTERNAL 66;
OPERATOR + ( A,B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 68;
OPERATOR + ( KA: INTEGER; B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 69;
OPERATOR + ( A: INTERVAL; KB: INTEGER ) RES: INTERVAL;
  EXTERNAL 70;
OPERATOR - ( A,B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 73;
OPERATOR - ( KA: INTEGER; B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 75;
OPERATOR - ( A: INTERVAL; KB: INTEGER ) RES: INTERVAL;
  EXTERNAL 74;
OPERATOR * ( A,B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 78;
OPERATOR * ( KA: INTEGER; B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 79;
OPERATOR * ( A: INTERVAL; KB: INTEGER ) RES: INTERVAL;
  EXTERNAL 80;
OPERATOR / ( A,B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 85;
OPERATOR / ( KA: INTEGER; B: INTERVAL ) RES: INTERVAL;
  EXTERNAL 83;
OPERATOR / ( A: INTERVAL; KB: INTEGER ) RES: INTERVAL;
  EXTERNAL 86;

FUNCTION ISCALP (VAR A, B: IVECTOR; AKDIM : INTEGER) : INTERVAL;
  EXTERNAL 88;

       (*  Standard Functions   *)

FUNCTION IABS ( Y: INTERVAL ) : REAL;
  EXTERNAL 101;
FUNCTION ISQR ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 102;
FUNCTION ISQRT ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 105;
FUNCTION IEXP ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 106;
FUNCTION ILN ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 107;
FUNCTION IARCTAN ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 108;
FUNCTION ISIN ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 109;
FUNCTION ICOS ( Y: INTERVAL ) : INTERVAL;
  EXTERNAL 110;
```

```
   (*  Input / Output  *)

PROCEDURE IREAD ( VAR F:TEXT; VAR A: INTERVAL );
 EXTERNAL 92;
PROCEDURE IWRITE ( VAR F: TEXT; A: INTERVAL );
 EXTERNAL 91;

FUNCTION ISUM (A: IVECTOR; DIM: DIMTYPE) : INTERVAL;
  VAR B: IVECTOR;
      I: DIMTYPE;
  BEGIN
  FOR I := 1 TO DIM DO B[I] := INTPT (1.0);
  ISUM := ISCALP (A, B, DIM)
  END;  (*  FUNCTION ISUM  *)

FUNCTION IVRNULL : IVECTOR;
  VAR I: DIMTYPE; U: IVECTOR;
  BEGIN
  FOR I := 1 TO DIM DO U[I] := INTPT (0.0);
  IVRNULL := U
  END;  (*  FUNCTION IVRNULL  *)

FUNCTION ITSQR (T: ITAYLOR) : ITAYLOR;                       (*  ITSQR(IT)  *)
  (*  Requires:  IVRNULL, ISCALP, ISQR  *)
  VAR I, J, K, HALF: DIMTYPE;
      X, Y: IVECTOR;
      U   : ITAYLOR;

  BEGIN
  X := IVRNULL;  Y := IVRNULL;
  U.LENGTH := T.LENGTH;
  U.T := T.T;
  U.TC := IVRNULL;

  U.TC[1] := ISQR (T.TC[1]);
  X[1] := T.TC[1];

  FOR K := 2 TO U.LENGTH DO
      BEGIN
      X[K] := T.TC[K];
      HALF := K DIV 2;
      FOR J := 1 TO HALF DO
          BEGIN
          I := K - J + 1;
          Y[J] := T.TC[I]
          END;  (* FOR J *)
      U.TC[K] := 2 * ISCALP ( X, Y, HALF);
      IF K MOD 2 = 1 THEN
          BEGIN
          HALF := HALF + 1;
          U.TC[K]:= U.TC[K] + ISQR (T.TC[HALF])
          END  (* IF *)
      END;  (* FOR K *)
  ITSQR := U
  END;  (* FUNCTION ITSQR (ITAYLOR)  *)
```

```
FUNCTION MENU_CHOICE : CHOICE;
    VAR I: INTEGER;
    BEGIN
    WRITELN;
    WRITELN ('ENTER:  1  -  GIVE NEW INITIAL CONDITIONS');
    WRITELN ('        2  -  CONTINUE EXTENDING THE SOLUTION');
    WRITELN ('        3  -  STOP');
    READ (I);
    IF ((I >= 3) OR (I <= 0)) THEN I := 3;
    MENU_CHOICE := I
    END;  (*  FUNCTION MENU_CHOICE  *)


PROCEDURE WRITE_INTERVAL (INT: INTERVAL);
    BEGIN
    WRITE ('[', INT.INF:12, ', ', INT.SUP:12, ']');
    END;  (*  PROCEDURE WRITE_INTERVAL  *)


PROCEDURE PRNT_ITAY_COEF (Y: ITAYLOR; INDEX: DIMTYPE);
    BEGIN
    WRITE ('Y(', INDEX:5, ') = ');
    WRITE_INTERVAL (Y.TC[INDEX]);
    WRITELN
    END;  (*  PROCEDURE PRNT_ITAY_COEF  *)


FUNCTION INTERVAL_LENGTH (INT: INTERVAL) : REAL;
    BEGIN
    INTERVAL_LENGTH := INT.SUP - INT.INF
    END;  (*  FUNCTION INTERVAL_LENGTH  *)


FUNCTION RELATIVE_LENGTH (INT: INTERVAL) : REAL;
    BEGIN
    RELATIVE_LENGTH := 2.0 * (INT.SUP - INT.INF)
                       / (INT.SUP + INT.INF)
    END;  (*  FUNCTION RELATIVE_LENGTH  *)


FUNCTION RELATIVE_ERROR (INT: INTERVAL) : REAL;
    BEGIN
    RELATIVE_ERROR := 2.0 * INT.SUP / (INT.SUP + INT.INF)
    END;  (*  FUNCTION RELATIVE_ERROR  *)



BEGIN  (*  MAIN PROGRAM IDEQ_SOLVE  *)

(*...........  INITIALIZE *)
FLAG := 2;
X.LENGTH := DIM;
Y.LENGTH := DIM;
RESET (DATA);
```

```
WHILE FLAG <= 2 DO              (* LOOP FOR NEW INITIAL CONDITIONS *)
    BEGIN
    FLAG := 2;
X.TC := IVRNULL;
Y.TC := IVRNULL;
    WRITELN ('READ INTERVAL INITIAL CONDITIONS X0, Y(X0):');
    IREAD (DATA, X.TC[1]); IREAD (DATA, Y.TC[1]);
    WRITELN;  WRITELN;
    WRITE ('INITIAL CONDITIONS AT X0 = ');
    WRITE_INTERVAL (X.TC[1]);  WRITELN (',');
    WRITE ('                      Y0 = ');
    WRITE_INTERVAL (Y.TC[1]);  WRITELN ('.');
    WHILE FLAG = 2 DO           (* LOOP FOR ANALYTIC CONTINUATION *)
        BEGIN
        (*..........  READ STEP SIZE *)
        WRITELN ('ENTER STEPSIZE X - X0:  ');
        READ (X.T);
        Y.T := X.T;
        WRITELN ('Computing series terms ...');
        FOR I := 2 TO DIM DO            (* LOOP FOR SERIES GENERATION *)
            BEGIN
(*      YOUR FIRST ORDER DIFFERENTIAL EQUATION GOES HERE:         *)
    YPRIME := ITSQR (Y);

            IM1 := I - 1;
            Y.TC[I] := YPRIME.TC[IM1] * INTPT (Y.T / IM1);
            END;  (*FOR*)

        (*..........  PRINT TABLE *)
        EPSILON := 0.5 * RELATIVE_LENGTH (Y.TC[1]);
        COMPOUND := 1.0;
        WRITELN;  WRITELN;
        WRITELN ('Step      Left         Right          Computed
Theoretical');
        WRITELN ('        Endpoint      Endpoint      Instability
Instability');
        WRITELN;
        FOR I := 1 TO DIM DO            (* LOOP FOR ERROR MEASUREMENT *)
            BEGIN
            COMPOUND := COMPOUND * (1.0 + EPSILON);
            WRITE (I:3);  WRITE ('  ');
            WRITE_INTERVAL (Y.TC[I]);
            WRITE ('   ', RELATIVE_ERROR (Y.TC[I]):10);
            WRITE ('     ', COMPOUND:10);  WRITELN
            END;  (*FOR*)
        (*..........  PERFORM THE ANALYTIC CONTINUATION *)
        Y.TC[1] := ISUM (Y.TC, DIM);
        FOR I :=2 TO DIM DO Y.TC[I] := INTPT (0.0);
        X.TC[1] := X.TC[1] + INTPT (X.T);
        WRITELN;
        WRITE ('THE VALUE AT X = ');  WRITE_INTERVAL (X.TC[1]);  WRITELN;
        WRITE ('          IS Y = ');  WRITE_INTERVAL (Y.TC[1]);  WRITELN ('.');
        FLAG := MENU_CHOICE
        END  (*WHILE*)
    END  (*WHILE*)
END.  (*MAIN PROGRAM IDEQ_SOLVE*)
```

APPENDIX C

PASCAL-SC REAL AND INTERVAL TAYLOR OPERATORS, PROCEDURES, AND FUNCTIONS

The operators, procedures, and functions are grouped into seven files. Source code can be found in the report [7].

```
1.  RIT_ADD.LIB  -  REAL AND INTERVAL TAYLOR ADD AND SUBTRACT    <<<<<<
2.  RIT_MUL.LIB  -  REAL AND INTERVAL TAYLOR MULTIPLY     <<<<<<<<<<<<<
3.  RIT_DIV.LIB  -  REAL AND INTERVAL TAYLOR DIVIDE     <<<<<<<<<<<<<<<
4.  RT_POW.LIB   -  REAL TAYLOR POWERS     <<<<<<<<<<<<<<<<<<<<<<<<<
5.  IT_POW.LIB   -  INTERVAL TAYLOR POWERS     <<<<<<<<<<<<<<<<<<<<<<
6.  RIT_FNS.LIB  -  REAL AND INTERVAL TAYLOR FUNCTIONS      <<<<<<<<<<<<
7.  UTIL.LIB     -  UTILITY PROCEDURES & FUNCTIONS     <<<<<<<<<<<<<<<<
```

The contents of each library are:

C.1. Addition and Subtraction Operators

```
RIT_ADD.LIB  -  REAL AND INTERVAL TAYLOR ADD AND SUBTRACT    <<<<<<
    + T         OPERATOR + (T: TAYLOR) RES : TAYLOR;
    K + T       OPERATOR + (K: INTEGER; T: TAYLOR) RES : TAYLOR;
    T + K       OPERATOR + (T: TAYLOR; K: INTEGER) RES : TAYLOR;
    R + T       OPERATOR + (R: REAL; T: TAYLOR) RES : TAYLOR;
    T + R       OPERATOR + (T: TAYLOR; R: REAL) RES : TAYLOR;
    T + T       OPERATOR + (TA, TB: TAYLOR) RES : TAYLOR;
    + IT        OPERATOR + (T: ITAYLOR) RES : ITAYLOR;
    K + IT      OPERATOR + (K: INTEGER; T: ITAYLOR) RES : ITAYLOR;
    IT + K      OPERATOR + (T: ITAYLOR; K: INTEGER) RES : ITAYLOR;
    I + IT      OPERATOR + (K: INTERVAL; T: ITAYLOR) RES : ITAYLOR;
    IT + I      OPERATOR + (T: ITAYLOR; K: INTERVAL) RES : ITAYLOR;
    IT + IT     OPERATOR + (TA, TB: ITAYLOR) RES : ITAYLOR;

    - T         OPERATOR - (T: TAYLOR) RES : TAYLOR;
    K - T       OPERATOR - (K: INTEGER; T: TAYLOR) RES : TAYLOR;
    T - K       OPERATOR - (T: TAYLOR; K: INTEGER) RES : TAYLOR;
    R - T       OPERATOR - (R: REAL; T: TAYLOR) RES : TAYLOR;
    T - R       OPERATOR - (T: TAYLOR; R: REAL) RES : TAYLOR;
    T - T       OPERATOR - (TA, TB: TAYLOR) RES : TAYLOR;
    - IT        OPERATOR - (T: ITAYLOR) RES : ITAYLOR;
    K - IT      OPERATOR - (K: INTEGER; T: ITAYLOR) RES : ITAYLOR;
    IT - K      OPERATOR - (T: ITAYLOR; K: INTEGER) RES : ITAYLOR;
    I - IT      OPERATOR - (K: INTERVAL; T: ITAYLOR) RES : ITAYLOR;
    IT - I      OPERATOR - (T: ITAYLOR; K: INTERVAL) RES : ITAYLOR;
    IT - IT     OPERATOR - (TA, TB: ITAYLOR) RES : ITAYLOR;
```

C.2. Multiplication Operators (Including TSQR and ITSQR)

```
RIT_MUL.LIB  -  REAL AND INTERVAL TAYLOR MULTIPLY     <<<<<<<<<<<<<
    K * T       OPERATOR * (K: INTEGER; T: TAYLOR) RES : TAYLOR;
    T * K       OPERATOR * (T: TAYLOR; K: INTEGER) RES : TAYLOR;
    R * T       OPERATOR * (R: REAL; T: TAYLOR) RES : TAYLOR;
    T * R       OPERATOR * (T: TAYLOR; R: REAL) RES : TAYLOR;
    T * T       OPERATOR * (TA, TB: TAYLOR) RES : TAYLOR;
    TSQR(T)     FUNCTION TSQR (T: TAYLOR) : TAYLOR;
```

C.2. Multiplication Operators (Continued)

```
    K * IT       OPERATOR * (K: INTEGER; T: ITAYLOR) RES : ITAYLOR;
    IT * K       OPERATOR * (T: ITAYLOR; K: INTEGER) RES : ITAYLOR;
    I * IT       OPERATOR * (K: INTERVAL; T: ITAYLOR) RES : ITAYLOR;
    IT * I       OPERATOR * (T: ITAYLOR; K: INTERVAL) RES : ITAYLOR;
    IT * IT      OPERATOR * (TA, TB: ITAYLOR) RES : ITAYLOR;
    ITSQR(IT)    FUNCTION ITSQR (T: ITAYLOR) : ITAYLOR;
```

C.3. Division Operators.

```
RIT_DIV.LIB  -   REAL AND INTERVAL TAYLOR DIVIDE     <<<<<<<<<<<<<<<
    K / T        OPERATOR / (K: INTEGER; T: TAYLOR) RES : TAYLOR;
    T / K        OPERATOR / (T: TAYLOR; K: INTEGER) RES : TAYLOR;
    R / T        OPERATOR / (R: REAL; T: TAYLOR) RES : TAYLOR;
    T / R        OPERATOR / (T: TAYLOR; R: REAL) RES : TAYLOR;
    T / T        OPERATOR / (TA, TB: TAYLOR) RES : TAYLOR;
    K / IT       OPERATOR / (K: INTEGER; T: ITAYLOR) RES : ITAYLOR;
    IT / K       OPERATOR / (T: ITAYLOR; K: INTEGER) RES : ITAYLOR;
    I / IT       OPERATOR / (K: INTERVAL; T: ITAYLOR) RES : ITAYLOR;
    IT / I       OPERATOR / (T: ITAYLOR; K: INTERVAL) RES : ITAYLOR;
    IT / IT      OPERATOR / (TA, TB: ITAYLOR) RES : ITAYLOR;
```

C.4. Real Power Operators and Functions.

```
RT_POW.LIB   -   REAL TAYLOR POWERS         <<<<<<<<<<<<<<<<<<<<<<<<<
    TSQR(T)      FUNCTION TSQR (T: TAYLOR) : TAYLOR;
    TSQRT(T)     FUNCTION TSQRT (T: TAYLOR) : TAYLOR;
    TEXP(T)      FUNCTION TEXP (T: TAYLOR) : TAYLOR;

    K ** K       OPERATOR ** (BASE, EXPONENT : INTEGER) RES : INTEGER;
    R ** K       OPERATOR ** (BASE: REAL; EXPONENT: INTEGER) RES : REAL;
    R ** R       OPERATOR ** (BASE, EXPONENT: REAL) RES : REAL;
    K ** R       OPERATOR ** (BASE: INTEGER; EXPONENT: REAL) RES : REAL;
    T ** K       OPERATOR ** (BASE: TAYLOR; EXPONENT: INTEGER) RES : TAYLOR;
    T ** R       OPERATOR ** (BASE: TAYLOR; EXPONENT: REAL) RES : TAYLOR;
    R ** T       OPERATOR ** (BASE: REAL; EXPONENT: TAYLOR) RES : TAYLOR;
    K ** T       OPERATOR ** (BASE: INTEGER; EXPONENT: TAYLOR) RES : TAYLOR;
    T ** T       OPERATOR ** (BASE, EXPONENT: TAYLOR) RES : TAYLOR;
```

C.5. Interval Power Operators and Functions.

```
IT_POW.LIB   -   INTERVAL TAYLOR POWERS       <<<<<<<<<<<<<<<<<<<<<<
    ITSQR(IT)    FUNCTION ITSQR (T: ITAYLOR) : ITAYLOR;
    ITSQR(IT)    FUNCTION ITSQR (T: ITAYLOR) : ITAYLOR;
    ITSQRT(IT)   FUNCTION ITSQRT (T: ITAYLOR) : ITAYLOR;
    ITEXP(IT)    FUNCTION ITEXP (T: ITAYLOR) : ITAYLOR;

    I ** K       OPERATOR ** (BASE: INTERVAL; EXPONENT: INTEGER) RES : INTERVAL;
    K ** I       OPERATOR ** (BASE: INTEGER; EXPONENT: INTERVAL) RES : INTERVAL;
    I ** I       OPERATOR ** (BASE: INTERVAL; EXPONENT: INTERVAL) RES : INTERVAL;
    IT ** K      OPERATOR ** (BASE: ITAYLOR; EXPONENT: INTEGER) RES : ITAYLOR;
    IT ** I      OPERATOR ** (BASE: ITAYLOR; EXPONENT: INTERVAL) RES : ITAYLOR;
    K ** IT      OPERATOR ** (BASE: INTEGER; EXPONENT: ITAYLOR) RES : ITAYLOR;
    I ** IT      OPERATOR ** (BASE: INTERVAL; EXPONENT: ITAYLOR) RES : ITAYLOR;
    IT ** IT     OPERATOR ** (BASE: ITAYLOR; EXPONENT: ITAYLOR) RES : ITAYLOR;
```

C.6. Real and Interval Functions and Procedures.

```
RIT_FNS.LIB   -   REAL AND INTERVAL TAYLOR FUNCTIONS      <<<<<<<<<<<<
    TSQR(T)       FUNCTION TSQR (T: TAYLOR) : TAYLOR;
    TSQRT(T)      FUNCTION TSQRT (T: TAYLOR) : TAYLOR;
    TEXP(T)       FUNCTION TEXP (T: TAYLOR) : TAYLOR;
    TLN(T)        FUNCTION TLN (T: TAYLOR) : TAYLOR;
    T_SIN_COS     PROCEDURE T_SIN_COS (T: TAYLOR; VAR S, C: TAYLOR);
    TSIN(T)       FUNCTION TSIN (T: TAYLOR) : TAYLOR;
    TCOS(T)       FUNCTION TCOS (T: TAYLOR) : TAYLOR;
    TRUNGE(T)     FUNCTION TRUNGE (T: TAYLOR) : TAYLOR;
    TARCTAN(T)    FUNCTION TARCTAN (T: TAYLOR) : TAYLOR;
    TTAN(T)       FUNCTION TTAN (T: TAYLOR) : TAYLOR;
    TDIFF(T)      FUNCTION TDIFF (T: TAYLOR) : TAYLOR;
    TINTGRL(T)    FUNCTION TINTGRL (T: TAYLOR) : TAYLOR;

    ITSQR(IT)     FUNCTION ITSQR (T: ITAYLOR) : ITAYLOR;
    ITSQRT(IT)    FUNCTION ITSQRT (T: ITAYLOR) : ITAYLOR;
    ITEXP(IT)     FUNCTION ITEXP (T: ITAYLOR) : ITAYLOR;
    ITLN(IT)      FUNCTION ITLN (T: ITAYLOR) : ITAYLOR;
    IT_SIN_COS    PROCEDURE IT_SIN_COS (T: ITAYLOR; VAR S, C: ITAYLOR);
    ITSIN(IT)     FUNCTION ITSIN (T: ITAYLOR) : ITAYLOR;
    ITCOS(IT)     FUNCTION ITCOS (T: ITAYLOR) : ITAYLOR;
    ITRUNGE(IT) FUNCTION ITRUNGE (T: ITAYLOR) : ITAYLOR;
    ITARCTAN(IT)  FUNCTION ITARCTAN (T: ITAYLOR) : ITAYLOR;
    ITTAN(IT)     FUNCTION ITTAN (T: ITAYLOR) : ITAYLOR;
    ITDIFF(T)     FUNCTION ITDIFF (T: ITAYLOR) : ITAYLOR;
    ITINTGRL(T) FUNCTION ITINTGRL (T: ITAYLOR) : ITAYLOR;
```

C.7. Utilities.

```
UTIL.LIB      -  UTILITY PROCEDURES & FUNCTIONS      <<<<<<<<<<<<<<<<
        FUNCTION VRNULL : RVECTOR;
        FUNCTION IVRNULL : IVECTOR;
        FUNCTION T_IDENT_ZERO (T: TAYLOR) : BOOLEAN;
        FUNCTION T_IDENT_CONSTANT (T: TAYLOR) : BOOLEAN;
        FUNCTION IT_IDENT_ZERO (T: ITAYLOR) : BOOLEAN;
        FUNCTION IT_IDENT_CONSTANT (T: ITAYLOR) : BOOLEAN;
        PROCEDURE WRITE_INTERVAL (INT: INTERVAL);
        PROCEDURE WRITE_INTERVAL_SERIES (SER: ITAYLOR);
        PROCEDURE READ_INTERVAL_SERIES (VAR F: ITAYLOR);
        PROCEDURE WRITE_SERIES (T: TAYLOR);
        FUNCTION ITMIDPT (F : ITAYLOR) : TAYLOR;
```

ON THE EXTREMUM OF BILINEAR FUNCTIONAL FOR HYPERBOLIC TYPE P.D.E.

C. N. Shen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. Transient solutions of the hyperbolic type partial differential equations are needed for solving many engineering problems such as computing stress waves for gun dynamics or determining shock behaviors of penetration mechanics.

Variational procedures using the bilinear formulations with adjoint variables can serve as the theoretical basis in the derivation of algorithms for the finite element methods, giving direct numerical solutions for partial derivatives of the functions to be found for these problems. The adjoint system can be arranged in a manner that it is a reflected mirror of the original system in time. Generalized boundary conditions employ many types of "springs" relating the various spatial partial derivatives. They are defined to satisfy the boundaries of the concomitant for the bilinear expression. Algorithms for use in the finite element method are simplified since the adjoint system gives exactly the same solutions as that of the original system. The second necessary condition for an extremum is satisfied by showing that the second variation is positive semi-definite.

I. INTRODUCTION. Transient solutions of the hyperbolic type partial differential equation, for example the wave equation or the beam equation, are important for solving engineering problems such as stress wave for gun dynamics or shock behavior of penetration mechanics. At present these equations are usually solved numerically by the finite difference method or by the Galerkin method. Considerable advantage may be obtained if the finite element method can be directly employed instead. Variational procedures using bilinear formulation with adjoint variables can serve as the theoretical basis for the derivation of algorithms using the finite element method for hyperbolic type p.d.e.

II. THE VARIATIONAL PRINCIPLE. A dynamical system can be modeled by the following partial differential equation.

$$L(\zeta)\ y(\zeta) = -Q(\zeta) \qquad (1)$$

with appropriate boundary and initial conditions. In the above equation L is a linear operator in both spatial and temporal domain, y is the dependent variable, Q is a forcing function, and ζ represents all independent variables, both spatial and temporal.

The inner produce < > of an adjoint forcing function $\bar{Q}$ and the solution $(y(\zeta))$ of Eq. (1) can be used for the purpose of estimation. This inner product is $\langle \bar{Q}, y \rangle$.

An accurate estimation can be made by constructing a variational principle [1]. By using the adjoint variable $\bar y$ as a Lagrange multiply for Eq. (1) adding to $\langle \bar Q, y\rangle$, we have

$$J_1[y,\bar y] \overset{\Delta}{=} \langle \bar Q, y\rangle + \langle \bar y,(Q+Ly)\rangle = \langle \bar Q, y\rangle + \langle \bar y, Q\rangle + \langle \bar y, Ly\rangle \qquad (2)$$

To keep the system symmetrical, let us define the adjoint system as

$$\bar L(\xi)\bar y(\xi) = -\bar Q(\xi) \qquad (3)$$

By using the original variable y as a Lagrange multiply for Eq. (3) adding to $\langle Q,\bar y\rangle$, we have

$$J_2[y,\bar y] \overset{\Delta}{=} \langle Q,\bar y\rangle + \langle y,(\bar Q+\bar L\bar y)\rangle = \langle Q,\bar y\rangle + \langle y,\bar Q\rangle + \langle y,\bar L\bar y\rangle \qquad (4)$$

By definition, the relationship of the adjoint system to the original system is

$$D \overset{\Delta}{=} \langle \bar y, Ly\rangle - \langle y,\bar L\bar y\rangle = 0 \qquad (5)$$

where D is the bilinear concomitant [1]. Combining Eqs. (2), (4), and (5) one obtains

$$J_1 = J_2 \qquad (6a)$$

In order to keep the functional symmetrical, we have

$$J \overset{\Delta}{=} \frac{1}{2}\,[J_1 + J_2] \qquad (6b)$$

which is of the form

$$J = \langle \bar Q, y\rangle + \langle \bar y, Q\rangle + \frac{1}{2}\langle \bar y, Ly\rangle + \frac{1}{2}\langle y,\bar L\bar y\rangle \qquad (6c)$$

To show that the above functional satisfies both the original and the adjoint systems, let us take the first variations of Eqs. (5) and (6) which gives

$$\delta J = \delta J(\delta\bar y) + \delta J(\delta y) \qquad (7a)$$

where

$$\delta J(\delta\bar y) = \langle \delta\bar y, Q\rangle + \frac{1}{2}\langle \delta\bar y, Ly\rangle + \frac{1}{2}\langle y,\bar L\delta\bar y\rangle = 0 \qquad (7b)$$

and

$$\delta J(\partial y) = \langle \delta y,\bar Q\rangle + \frac{1}{2}\langle \delta y,\bar L\bar y\rangle + \frac{1}{2}\langle \bar y, L\delta y\rangle = 0 \qquad (7c)$$

Also

$$\delta D = \delta D(\delta\bar y) + \delta D(\delta y) \qquad (8a)$$

where

$$\delta D(\delta\bar{y}) = \langle\delta\bar{y},Ly\rangle - \langle y,\bar{L}\delta\bar{y}\rangle = 0 \qquad (8b)$$

and

$$\delta D(\delta y) = - \langle\delta y,\overline{Ly}\rangle + \langle\bar{y},L\delta y\rangle = 0 \qquad (8c)$$

From Eqs. (7b) and (8b) we obtained

$$\delta J(\delta\bar{y}) = \langle\delta\bar{y},Q) + \frac{1}{2}\langle\delta\bar{y},Ly\rangle + \frac{1}{2}\langle\delta\bar{y},Ly\rangle = \langle\delta\bar{y},(Q+Ly)\rangle = 0 \qquad (9)$$

For arbitrary $\delta\bar{y}$ satisfying certain general limitations on the boundaries it can be shown that the Euler-Lagrange Equation for the original system in Eq. (1) is satisfied. From Eqs. (7c) and (8c) we get

$$\delta J(\delta y) = \langle\delta y,\bar{Q}\rangle - \frac{1}{2}\langle\delta y,\overline{Ly}\rangle + \frac{1}{2}\langle\delta y,\overline{Ly}\rangle = (\delta y,(\bar{Q}+\overline{Ly}))\rangle = 0 \qquad (10)$$

For arbitrary variation δy, the Euler-Lagrange Equation for the adjoint system in Eq. (3) is also satisfied.

III. INTEGRAL OF BILINEAR EXPRESSION. The integral of a bilinear expression for a two dimensional problem having second order partial derivatives in time and fourth order partial derivatives in space can be written as

$$I = \int_{x_o}^{x_b}\int_{t_o}^{t_b} \Omega[y(x,t),\bar{y}(x,t)]dtdx \qquad (11)$$

where $\Omega[y,\bar{y}]$ is a given bilinear expression in the form

$$\Omega[y,\bar{y}] = y_t\bar{y}_t - \omega^2 y\bar{y} - a^2 y_x\bar{y}_x - b^2 y_{xx}\bar{y}_{xx} \qquad (12)$$

The subscripts t and x indicate the partial derivatives for the functions y and $\bar{y}$.

Equation (12) can be integrated by parts. Two different forms of integration and end conditions are given. The first form of the integral is obtained by integrating by parts on the adjoint variable.

$$I_1 = -\int_{t_o}^{t_b}\int_{x_o}^{x_b} \bar{y}Lydtdx + \int_{x_o}^{x_b} y_t\bar{y}\Big|_{t_o}^{t_b}dx$$

$$+ \int_{t_o}^{t_b} \{-b^2 y_{xx}\bar{y}_x\Big|_{x_o}^{x_b} + (b^2 y_{xx})_x\bar{y}\Big|_{x_o}^{x_b} - a^2 y_x\bar{y}\Big|_{x_o}^{x_b}\}dt \qquad (13a)$$

On the other hand, we can perform integration on the original variable to give

$$I_2 = -\int_{t_o}^{t_b}\int_{x_o}^{x_b} y\overline{Ly}\,dtdx + \int_{x_o}^{x_b} \bar{y}_t y\Big|_{t_o}^{t_b} dx$$

$$+ \int_{t_o}^{t_b} \{-b^2\bar{y}_{xx}y_x\Big|_{x_o}^{x_b} + b^2(\bar{y}_{xx})_x y\Big|_{x_o}^{x_b} - a^2\bar{y}_x y\Big|_{x_o}^{x_b}\}dt \tag{13b}$$

To keep the form symmetrical, we take the average of the above two expressions

$$I = \frac{1}{2}I_1 + \frac{1}{2}I_2 = -\int_{x_o}^{x_b}\int_{t_o}^{t_b} \frac{1}{2}(\bar{y}Ly+y\overline{Ly})dtdx + \frac{1}{2}\int_{x_o}^{x_b}(y_t\bar{y}+\bar{y}_t y)\Big|_{t_o}^{t_b} dx$$

$$+ \frac{1}{2}\int_{t_o}^{t_b}(-a^2)(y_x\bar{y}+\bar{y}_x y)\Big|_{x_o}^{x_b} dt$$

$$+ \frac{1}{2}\int_{t_o}^{t_b}(-b^2)(y_{xx}\bar{y}_x+\bar{y}_{xx}y_x)\Big|_{x_o}^{x_b} dt - \frac{1}{2}\int_{t_o}^{t_b}[(-b^2y_{xx})_x\bar{y} + (-b^2\bar{y}_{xx})_x y]_{x_o}^{x_b} dt \tag{14}$$

where

$$Ly = y_{tt} + \omega^2 y - a^2 y_{xx} + b^2 y_{xxxx} \tag{15a}$$

and

$$\overline{Ly} = \bar{y}_{tt} + \omega^2\bar{y} - a^2\bar{y}_{xx} + b^2\bar{y}_{xxxx} \tag{15b}$$

For a fourth order spatial partial and a second order temporal partial system Eq. (5) becomes

$$D = \int_{x_o}^{x_b}\int_{t_o}^{t_b} \bar{y}Ly\,dtdx - \int_{x_o}^{x_b}\int_{t_o}^{t_b} y\overline{Ly}\,dtdx \tag{16a}$$

By equating Eqs. (13a) and (13b) and solving for D in Eq. (16a) we are converting the double integral into single integrals in terms of the boundary conditions.

We can express the quantity D as the sum of three parts for end conditions D_1, D_2, and D_3 as

$$D = D_1 + D_2 + D_3 \tag{16b}$$

The terms in D_1 involve the initial conditions of y and $\bar{y}$ as

$$D_1 = \int_{x_o}^{x_b} \{y_t\bar{y}\Big|_{t_o}^{t_b} - \bar{y}_t y\Big|_{t_o}^{t_b}\}dx$$

$$D_1 = \int_{x_o}^{x_b} dx\{[y_t(x,t_b)\bar{y}(x,t_b) - \bar{y}_t(x,t_b)y(x,t_b)]$$

$$- [\bar{y}_t(x,t_o)\bar{y}(x,t_o) - \bar{y}_t(x,t_o)y(x,t_o)]\} \qquad (17a)$$

The terms in D_2 involve the boundary conditions from the second partials of y and $\bar{y}$ as

$$D_2 = \int_{t_o}^{t_b} (-a^2)\{y_x\bar{y}\Big|_{x_o}^{x_b} - \bar{y}_x y\Big|_{x_o}^{x_b}\}dt$$

$$D_2 = \int_{t_o}^{t_b} dt\{-a^2[y_x(x_b,t)\bar{y}(x_b,t) - \bar{y}_x(x_b,t)y(x_b,t)$$

$$+ a^2[y_x(x_o,t)\bar{y}(x_o,t) - \bar{y}_x(x_o,t)y(x_o,t)]\} \qquad (17b)$$

The terms in D_3 involve the boundary conditions from the fourth partials of y and $\bar{y}$ as

$$D_3 = \int_{t_o}^{t_b} \{-b^2y_{xx}\bar{y}_x\Big|_{x_o}^{x_b} + b^2\bar{y}_{xx}y_x\Big|_{x_o}^{x_b} + (b^2y_{xx})_x\bar{y}\Big|_{x_o}^{x_b} + (-b^2\bar{y}_{xx})_xy\Big|_{x_o}^{x_b}\}dt$$

$$D_3 = \int_{t_o}^{t_b} dt\{-b^2[y_{xx}(x_b,t)\bar{y}_x(x_b,t) - \bar{y}_{xx}(x_b,t)y_x(x_b,t)]$$

$$+ b^2[y_{xx}(x_o,t)\bar{y}_x(x_o,t) - \bar{y}_{xx}(x_o,t)y_x(x_o,t)]\}$$

$$+ \int_{t_o}^{t_b} dt\{-b^2[-y_{xxx}(x_b,t)\bar{y}(x_b,t) + \bar{y}_{xxx}(x_b,t)y(x_b,t)]$$

$$+ b^2[-y_{xxx}(x_o,t)\bar{y}(x_o,t) + \bar{y}_x(x_o,t)y(x_o,t)]\} \qquad (17c)$$

In order that $D \equiv 0$ in Eq. (16b) it is sufficient that

$$D_1 \equiv 0 \qquad (18a)$$

$$D_2 \equiv 0 \qquad (18b)$$

and

$$D_3 \equiv 0 \qquad (18c)$$

IV. THE SYMMETRICAL ADJOINT SYSTEM. The adjoint independent variable τ in Figure 1 can be expressed as

$$\frac{\tau_b - \tau}{\tau_b - \tau_o} = \frac{t - t_o}{t_b - t_o} \qquad (19)$$

which gives

$$\tau = \tau_b \quad \text{for} \quad t = t_o \tag{20a}$$

and

$$\tau = \tau_o \quad \text{for} \quad t = t_b \tag{20b}$$

It is noted from Eq. (19) that

$$\tau_b - \tau_o = t_b - t_o \tag{21a}$$

$$\tau = \tau_b + t_o - t \tag{21b}$$

$$d\tau = -dt \tag{21c}$$

$$\frac{d}{d\tau} = - \frac{d}{dt} \tag{21d}$$

and

$$\bar{y}(x,t) = y(x,\tau = \tau_b+t_o-t) \tag{21e}$$

Let us assume that the adjoint system shown in Figure 1 gives

$$\bar{y}(x,\hat{t}=t) = y(x,\hat{t}=t_b+t_o-t) \tag{22a}$$

$$\bar{y}_t(x,\hat{t}=t) = -y_t(x,\hat{t}=t_b+t_o-t) \tag{22b}$$

$$\bar{y}_x(x,\hat{t}=t) = y_x(x,\hat{t}=t_b+t_o-t) \tag{22c}$$

where $\hat{t}$ is a dummy variable for t.

We may define the adjoint system as the image reflection in the time domain of the original system. Equation (22) yields the following known initial conditions

$$\bar{y}(x,\hat{t}=t_b) = y(x,\hat{t}=t_o) \quad \text{(known)} \tag{23a}$$

$$\bar{y}_t(x,\hat{t}=t_b) = -y_t(x,\hat{t}=t_o) \quad \text{(known)} \tag{23b}$$

The interpretation of the above equations gives the initial conditions of the original system as the far end conditions for the adjoint system, since the adjoint system is a reflected mirror of the original system in time.

V. INITIAL CONDITIONS FOR THE ADJOINT SYSTEM. We take a symmetry approach for the initial conditions of the adjoint system as

$$\bar{y}(x,t=t_b) = y(x,t=t_o) \quad , \quad \bar{y}_t(x,t=t_b) = -y_t(x,t=t_o) \tag{24}$$

$$\bar{y}(x,t=t_o) = y(x,t=t_b) \quad , \quad \bar{y}_t(x,t=t_o) = -y_t(x,t=t_b) \tag{25}$$

Thus Eq. (17a) becomes

$$D_1 = \int_{x_o}^{x_b} dx\{[y_t(x,t=t_b)y(x,t=t_o) + y_t(x,t=t_o)y(x,t=t_b)$$

$$- [y_t(x,t=t_o)y(x,t=t_b) + y_t(x,t=t_b)y(x,t=t_o)]\} = 0 \tag{26}$$

Since the integrand of Eq. (26) is zero, the above satisfies Eq. (18a). The initial conditions in Eq. (25) are given. Therefore

$$\delta\bar{y}(x,t=t_b) = \delta y(x,t=t_o) = 0 \tag{27a}$$

$$\delta\bar{y}_t(x,t=t_b) = -\delta y_t(x,t=t_o) = 0 \tag{27b}$$

VI. THE GENERALIZED BOUNDARY CONDITIONS. Let us consider the operator L in Eq. (15a) for two different cases as follows.

A. For the wave equation we have

$$Ly = y_{tt} - a^2 y_{xx} \tag{28}$$

Let us assume that elastic springs are installed at the ends such that

$$y_x(x_b,t) = k_b y(x_b,t) \quad , \quad \bar{y}_x(x_b,t) = k_b \bar{y}(x_b,t) \tag{29a}$$

$$y_x(x_o,t) = -k_o y(x_o,t) \quad , \quad \bar{y}_x(x_o,t) = -k_o \bar{y}(x_o,t) \tag{29b}$$

This is equivalent to state that the fixed end condition for a prismatic bar is $k_b = k_o \to \infty$ and the free end condition is $k_b = k_o \to 0$. If Eq. (29) is substituted into Eq. (17b) we have

$$D_2 = 0 \tag{30}$$

B. For the beam equation we have

$$Ly = y_{tt} + b^2 y_{xxxx} \tag{31}$$

Two sets of springs are incorporated at the ends. They are:

(1) Torsional springs relates the moments (the second partials) with the slopes (the first partials)

$$y_{xx}(x_b,t) = n_b y_x(x_b,t) \qquad \bar{y}_{xx}(x_b,t) = n_b \bar{y}_x(x_b,t) \tag{32a}$$

$$y_{xx}(x_o,t) = -n_o y_x(x_o,t) \qquad \bar{y}_{xx}(x_o,t) = -n_o \bar{y}_x(x_o,t) \tag{32b}$$

(2) Linear springs relates the shears (the third partials) with the deflection (no partials)

$$y_{xxx}(x_b,t) = C_b y(x_b,t) \qquad \bar{y}_{xxx}(x_b,t) = C_b \bar{y}(x_b,t) \tag{33a}$$

$$y_{xxx}(x_o,t) = -C_o y(x_o,t) \qquad \bar{y}_{xxx}(x_o,t) = -C_o \bar{y}(x_o,t) \tag{33b}$$

By substituting Eqs. (32) and (33) into Eq. (17c) we have

$$D_3 = 0 \tag{34}$$

Table I shows the assignment of the spring constants for various physical end conditions.

TABLE I. GENERALIZED BOUNDARY CONDITIONS

	At Fixed End	At Hinged End	At Guided End	At Free End
	$y=\bar{y}=0$ $y_x=\bar{y}_x=0$ $\delta y=\delta\bar{y}=0$ $\delta y_x=\delta\bar{y}_x=0$	$y=\bar{y}=0$ $y_{xx}=\bar{y}_{xx}=0$ $\delta y=\delta\bar{y}=0$ $\delta y_{xx}=\delta\bar{y}_{xx}=0$	$y_x=\bar{y}_x=0$ $y_{xxx}=\bar{y}_{xxx}=0$ $\delta y_x=\delta\bar{y}_x=0$ $\delta y_{xxx}=\delta\bar{y}_{xxx}=0$	$y_{xx}=\bar{y}_{xx}=0$ $y_{xxx}=\bar{y}_{xxx}=0$ $\delta y_{xx}=\delta\bar{y}_{xx}=0$ $\delta y_{xxx}=\delta\bar{y}_{xxx}=0$
Torsional Spring $\overset{\Delta}{y}_{xx}=\eta y_x$	$\eta \to \infty$	$\eta \to 0$	$\eta \to \infty$	$\eta \to 0$
Deflection Spring $y_{xxx}=cy$	$c \to \infty$	$c \to \infty$	$c \to 0$	$c \to 0$
Spring $y_x=ky$	$\delta y=\delta\bar{y}=0$ $\delta y_x=\delta\bar{y}_x=0$	$k \to \infty$	$k \to 0$	k = undetermined

VII. THE FIRST VARIATION. The sum of the two functionals is obtained by adding Eqs. (6c) and (14) as

$$J + I = \int_{x_o}^{x_b} \int_{t_o}^{t_b} (Q\bar{y}+y\bar{Q})dxdt + T + W + B \tag{35}$$

where

$$T = \frac{1}{2} \int_{x_o}^{x_b} (y_t\bar{y}+\bar{y}_t y)\Big|_{t_o}^{t_b} dx \quad , \quad W = \frac{1}{2} \int_{t_o}^{t_b} (-a^2)(y_x\bar{y}+\bar{y}_x y)\Big|_{x_o}^{x_b} dt$$

and

$$B = \frac{1}{2} \int_{t_o}^{t_b} (-b^2)(y_{xx}\bar{y}_x+\bar{y}_{xx}y_x)\Big|_{x_o}^{x_b} dt - \frac{1}{2} \int_{t_o}^{t_b} [(-b^2y_{xx})_x\bar{y} + (-b^2\bar{y}_{xx})_x y\Big|_{x_o}^{x_b} dt \tag{36}$$

By taking the variations $\delta\bar{y}$ and δy separately, we let

$$\delta J = \delta J(\delta\bar{y}) + \delta J(\delta y) \tag{37}$$

Then one obtains from Eqs. (35) and (36) that

$$\delta J(\delta\bar{y}) = -\delta I(\delta\bar{y}) + \iint Q\delta\bar{y}\,dxdt + \delta I(\delta\bar{y}) + \delta W(\delta\bar{y}) + \delta B(\delta\bar{y}) = 0$$

where

$$\delta T(\delta\bar{y}) = \frac{1}{2}\int_{x_o}^{x_b} (y_t\delta\bar{y}+y\delta\bar{y}_t)\Big|_{t_o}^{t_b} dx \quad , \quad \delta W(\delta\bar{y}) = \frac{1}{2}\int_{t_o}^{t_b} (-a^2)(y_x\delta\bar{y}+y\delta\bar{y}_x)\Big|_{x_o}^{x_b} dt$$

and

$$\delta B(\delta\bar{y}) = \frac{1}{2}\int_{t_o}^{t_b} (-b^2)(y_{xx}\delta\bar{y}_x+y_x\delta\bar{y}_{xx})\Big|_{x_o}^{x_b} dt$$

$$- \frac{1}{2}\int_{t_o}^{t_b} [(-b^2)y_{xxx}\delta\bar{y} + (-b^2)y\delta\bar{y}_{xxx}]_{x_o}^{x_b}\, dt \tag{38}$$

where $-\delta I(\delta\bar{y})$ can be derived from Eqs. (11) and (12) as

$$-\delta I(\delta\bar{y}) = -\int_{x_o}^{x_b}\int_{t_o}^{t_b} (y_t\delta\bar{y}_t-\omega^2y\delta\bar{y}-a^2y_x\delta\bar{y}_x-b^2y_{xx}\delta\bar{y}_{xx})dxdt \tag{39}$$

The second term on the right side of Eq. (37) is

$$\delta J(\delta y) = -\delta I(\delta y) + \iint \bar{Q}\delta y\,dxdt + \delta T(\delta y) + \delta W(\delta y) + \delta B(\delta y) = 0$$

where

$$\delta T(\delta y) = \frac{1}{2}\int_{x_o}^{x_b} (\bar{y}_t\delta y+\bar{y}\delta y_t)\Big|_{t_o}^{t_b} dx \quad , \quad \delta W(\delta y) = \frac{1}{2}\int_{t_o}^{t_b} (-a^2)(\bar{y}_x\delta y+\bar{y}\delta y_x)\Big|_{x_o}^{x_b} dt$$

$$\delta B(\delta y) = \frac{1}{2}\int_{t_o}^{t_b} (-b^2)(\bar{y}_{xx}\delta y_x+\bar{y}_x\delta y_{xx})\Big|_{x_o}^{x_b} dt$$

$$- \int_{t_o}^{t_b} [(-b^2\bar{y}_{xxx})\delta y-b^2\bar{y}\delta y_{xxx}]_{x_o}^{x_b} dt \tag{40}$$

and

$$-\delta I(\delta y) = -\int_{x_o}^{x_b}\int_{t_o}^{t_b} (\bar{y}_t\delta y_t-\omega^2\bar{y}\delta y-a^2\bar{y}_x\delta y_x(-b^2)\bar{y}_{xx}\delta y_{xx})dxdt \tag{41}$$

It is noted that Eqs. (38) and (40) are exactly the same form, where Eqs. (39) and (41) are also similar.

For the beam equation it is noted that the high partials in Eqs. (38) and (39) can be replaced by Eqs. (32) and (33). The variations of the adjoint higher partials from these equations can be written as

$$\delta\bar{y}_{xx}(x_b,t) = n_b\delta\bar{y}_x(x_b,t) \qquad \delta\bar{y}_{xxx}(x_b,t) = c_b\delta\bar{y}(x_b,t) \tag{42a}$$

$$\delta\bar{y}_{xx}(x_o,t) = -n_o\delta\bar{y}_x(x_o,t) \qquad \delta\bar{y}_{xxx}(x_o,t) = -c_o\delta\bar{y}(x_o,t) \tag{42b}$$

Equations (38) and (39), with the aid of Eqs. (32), (33), and (42), are the key equations to be used for the finite element method. It is noted that the first variation $\delta J(\delta\bar{y})$ is the same as the first variation $\delta J(\delta y)$ by adding or dropping the bar on top of the variables and their variations. We do not need to solve for the adjoint system in Eqs. (40) and (41) since they give exactly the same solutions as that of the original system.

VIII. SECOND VARIATIONS. The functions y and $\bar{y}$ and their partials are written in the form in terms of a small parameter μ

$$y(x,t,\mu) = y(x,t) + \delta y(x,t,\mu) \quad , \quad \delta y(x,t,u) = \mu\eta(x,t) \tag{43a}$$

$$y_t(x,t,\mu) = y_t(x,t) + \delta y_t(x,t,\mu) \quad , \quad \delta y_t(x,t,u) = \mu\eta_t(x,t) \tag{43b}$$

$$y_x(x,t,\mu) = y_x(x,t) + \delta y_x(x,t,\mu) \quad , \quad \delta y_x(x,t,u) = \mu\eta_x(x,t) \tag{43c}$$

$$\bar{y}(x,t,\mu) = \bar{y}(x,t) + \delta\bar{y}(x,t,\mu) \quad , \quad \delta\bar{y}(x,t,u) = \mu\bar{\eta}(x,t) \tag{43d}$$

$$\bar{y}_t(x,t,\mu) = \bar{y}_t(x,t) + \delta\bar{y}_t(x,t,\mu) \quad , \quad \delta\bar{y}_t(x,t,u) = \mu\bar{\eta}_t(x,t) \tag{43e}$$

$$\bar{y}_x(x,t,\mu) = \bar{y}_x(x,t) + \delta\bar{y}_x(x,t,\mu) \quad , \quad \delta\bar{y}_x(x,t,u) = \mu\bar{\eta}_x(x,t) \tag{43f}$$

Similar expressions can be derived for higher partials in x. Thus, the functional $J(\mu)$ can be expressed [2] as

$$J(u) = J(u=0) + \delta J + \delta^2 J \tag{44a}$$

where

$$\delta J = \mu\left(\frac{\partial J}{\partial \mu}\right)_{u=0} \tag{44b}$$

and

$$\delta^2 J = \frac{\mu^2}{2}\left(\frac{\partial^2 J}{\partial \mu^2}\right)_{u=0} \tag{44c}$$

By taking variations of $\delta J(\delta\bar{y})$ in Eqs. (38) and (39) and some for $\delta J(\delta y)$ in Eqs. (40) and (41), we have

$$\delta^2 J = \delta^2 T + \delta^2 B + \delta^2 W - \delta^2 I \tag{45a}$$

where

$$\delta^2 T = \frac{1}{2}\int_{x_b}^{x_o} (\delta y_t\delta\bar{y} + \delta y\delta\bar{y}_t)\Big|_{t_o}^{t_b} dx \tag{45b}$$

$$\delta^2 B = \frac{1}{2}\int_{t_o}^{t_b} (-b^2)(\delta y_{xx}\delta\bar{y}_x+\delta y_x\delta\bar{y}_{xx})\Big|_{x_o}^{x_b} dt$$

$$+ \frac{1}{2}\int_{t_o}^{t_b} b^2(\delta y_{xxx}\delta\bar{y}+\delta y\delta\bar{y}_{xxx})\Big|_{x_o}^{x_b} dt \qquad (45c)$$

and

$$\delta^2 W = \frac{1}{2}\int_{t_o}^{t_b} (-a^2)(\delta y_x\delta\bar{y}+\delta y\delta\bar{y}_x)\Big|_{x_o}^{x_b} dt \qquad (45d)$$

The second variation of I is obtained from Eqs. (39) and (41) as

$$\delta^2 I = \frac{\mu^2}{2}\left(\frac{\partial^2 I}{\partial\mu^2}\right)_{\mu=0}$$

$$= \frac{1}{2}\delta y[\delta I(\delta y)] + \frac{1}{2}\delta\bar{y}[\delta I(\delta y)]$$

$$= \frac{1}{2}\int_{x_o}^{x_b}\int_{t_o}^{t_b} (\delta y_t\delta\bar{y}_t-\omega^2\delta y\delta\bar{y}-a^2\delta y_x\delta\bar{y}_x-b^2\delta y_{xx}\delta\bar{y}_{xx})dxdt$$

$$+ \frac{1}{2}\int_{x_o}^{x_b}\int_{t_o}^{t_b} (\delta\bar{y}_t\delta y_t-\omega^2\delta\bar{y}\delta y-a^2\delta\bar{y}_x\delta y_x-b^2\delta\bar{y}_{xx}\delta y_{xx})dxdt$$

$$\delta^2 I = \int_{x_o}^{x_b}\int_{t_o}^{t_b} (\delta\bar{y}_t\delta y_t-\omega^2\delta\bar{y}\delta y-a^2\delta\bar{y}_x\delta y_x-b^2\delta\bar{y}_{xx}\delta y_{xx})dxdt \qquad (45e)$$

Substituting Eq. (27) into Eq. (45b) we have

$$\delta^2 T = 0 \qquad (46a)$$

For all the end conditions in Table I either the variations δy_{xx} and $\delta\bar{y}_{xx}$ must vanish or δy_x and δy_x must vanish. Thus, the first term on the right side of Eq. (45c) is zero. Similarly, for all the end conditions in Table I either the variations δy_{xx} and δy_{xx} must vanish or δy and δy must vanish. Thus the second term on the right side of Eq. (45c) is also zero. The third term is zero except at the guided end. Thus, in general

$$\delta^2 B = 0 \qquad (46b)$$

In Table I, except the free end, either the δy_x and $\delta\bar{y}_x$ must vanish or δy and δy must vanish. Thus, one obtains

$$\delta^2 W = 0 \qquad (46c)$$

This reduces the second variations $\delta^2 J$ to

$$\delta^2 J = -\delta^2 \mathbf{I} \tag{47}$$

as given in Eq. (45e).

Substituting Eq. (45e) into Eq. (47) gives

$$\delta^2 J = \int_{t_o}^{t_b} \int_{x_o}^{x_b} [-\delta y_t(x,t)\delta\bar{y}_t(x,t) + \omega^2 \delta y(x,t)\delta\bar{y}(x,t) + $$

$$+ a^2 \delta y_x(x,t)\delta\bar{y}_x(x,t) + b^2 \delta y_{xx}(x,t)\delta\bar{y}_{xx}(x,t)]dxdt \tag{48}$$

In order that the functional J is an extremum [3,4], the second variation $\delta^2 J$ must be either positive semi-definite or negative semi-finite, i.e.,

$$\delta^2 J > 0 \quad (\text{or } \delta^2 J < 0) \tag{49}$$

The above is a necessary condition for a minimum (or a maximum).

The adjoint variations in Eq. (48) may be obtained by the relations given in Eq. (22) as

$$\delta\bar{y}(x,\hat{t}=t) = \delta y(x,\hat{t}=t_b+t_o-t) \tag{50a}$$

$$\delta\bar{y}_t(x,\hat{t}=t) = -\delta y_t(x,\hat{t}=t_b+t_o-t) \tag{50b}$$

$$\delta\bar{y}_x(x,\hat{t}=t) = \delta y_x(x,\hat{t}=t_b+t_o-t) \tag{50c}$$

The variations of adjoint initial conditions can be derived from Eq. (23) as

$$\delta\bar{y}(x,\hat{t}=t_b) = \delta y(x,\hat{t}=t_b) = 0 \quad \text{for all } x \tag{51a}$$

$$\delta\bar{y}_t(x,\hat{t}=t_b) = -\delta y_t(x,t=t_o) = 0 \quad \text{for all } x \tag{51b}$$

By substituting Eq. (51) into Eq. (48), we have

$$\delta^2 J = \int_{t_o}^{t_b} \int_{x_o}^{x_b} P(x,t)dxdt \tag{52a}$$

where

$$P(x,t) = \delta y_t(x,t)\delta y_t(x,t_b+t_o-t) + \omega^2 \delta y(x,t)\delta y_x(x,t_b+t_o-t)$$

$$+ a^2 \delta y_x(x,t)\delta y_x(x,t_b+t_o-t) + b^2 \delta y_{xx}(x,t)\delta y_{xx}(x,t_b+t_o-t) \tag{52b}$$

IX. SENSITIVITY RELATIONSHIP. In order to show that the second variation of the functional J is positive semi-definite, one needs to obtain the variations of the function and its partials together with that of the adjoint functions and its partials as indicated in Eq. (48). We can get these variations through the study of the sensitivity coefficients [5] and its relationship to the parameters given in Eq. (43). Let the forcing function in Eq. (1) be

$$Q(x,t) = qf(x,t) \tag{53}$$

It is assumed that the forcing function parameter q is subject to a small constant perturbation δq as

$$q = q_o + \delta q \tag{54}$$

Then the variation of the function y is

$$\delta y(x,t) = \frac{\partial y(x,t)}{\partial q} \delta q = \nu(x,t)\delta q \tag{55a}$$

where

$$\nu(x,t) = \frac{\partial y}{\partial q} \tag{55b}$$

The quantity ν is the sensitivity coefficient for the variation $\delta y(x,t)$ due to a small constant perturbation δq.

The original p.d.e. in Eq. (15a) can be written as

$$\begin{aligned}\phi &= Ly + Q \\ &= y_{tt} + \omega^2 y - a^2 y_{xx} + b^2 y_{xxxx} + qf(x,t) = 0 \end{aligned}\tag{56}$$

Due to the perturbation of q the change of ϕ obeys the following relationship

$$\frac{\partial\phi}{\partial y_{tt}} \frac{\partial y_{tt}}{\partial q} + \omega^2 \frac{\partial y}{\partial q} + \frac{\partial\phi}{\partial y_{xx}} \frac{\partial y_{xx}}{\partial q} + \frac{\partial\phi}{\partial y_{xxxx}} \frac{\partial y_{xxxx}}{\partial q} + f(x,t) = 0 \tag{57}$$

It is also noted from Eq. (56) that

$$\frac{\partial\phi}{\partial y_{tt}} = 1 \quad , \quad \frac{\partial\phi}{\partial y} = \omega^2 \tag{58a}$$

$$\frac{\partial\phi}{\partial y_{xx}} = -a^2 \quad , \quad \frac{\partial\phi}{\partial y_{xxxx}} = b^2 \tag{58b}$$

Using the definition in Eq. (55b) the partials can be interchanged as

$$\frac{\partial y_{tt}}{\partial q} = \frac{\partial^2}{\partial t^2}\left(\frac{\partial y}{\partial q}\right) = \nu_{tt} \tag{59a}$$

$$\frac{\partial y_{xx}}{\partial q} = \frac{\partial^2}{\partial x^2}\left(\frac{\partial y}{\partial q}\right) = \nu_{xx} \tag{59b}$$

and

$$\frac{\partial y_{xxxx}}{\partial q} = \frac{\partial^4}{\partial x^4}\left(\frac{\partial y}{\partial q}\right) = \nu_{xxxx} \tag{59c}$$

Substituting Eqs. (58) and (59) into Eq. (57) we have

$$\nu_{tt} + \omega^2\nu - a^2\nu_{xx} + b^2\nu_{xxxx} + f(x,t) = 0 \tag{60}$$

If we compare the definitions of variation in Eq. (43a) with the definition of sensitivity relationship in Eq. (55a) we have

$$\delta y(x,t) = \mu\eta(x,t) = (\delta q)\nu(x,t) \tag{61}$$

which gives

$$\eta(x,t) = \nu(x,t) \tag{62a}$$

and

$$\delta q = \mu \tag{62b}$$

Thus Eq. (60) becomes

$$\eta_{tt} + \omega^2\eta - a^2\eta_{xx} + b^2\eta_{xxxx} + f(x,t) = 0 \tag{63}$$

which gives the p.d.e. of the variations of the original system.

If we compare Eq. (63) with Eq. (56) we see that the variation $\eta(x,t) = \mu^{-1}\delta y(x,t)$ in Eq. (63) takes the place of the function y in Eq. (56) with q = 1. Therefore, the p.d.e. for the variations is unchanged except by a scale factor. Thus the solution of the variation $\delta y(x,t)$ has the same form as that of the original function y.

Similarly for the adjoint system one can obtain

$$\delta\bar{y}(x,t) = \overline{\mu\eta}(x,t) = (\delta\bar{q})\bar{\nu}(x,t) \tag{64}$$

$$\bar{\eta}(x,t) = \bar{\nu}(x,t) \tag{65a}$$

$$\delta\bar{q} = \bar{\mu} \tag{65b}$$

and

$$\bar{\eta}_{tt} + \omega^2\bar{\eta} - a^2\bar{\eta}_{xx} + b^2\bar{\eta}_{xxxx} + \bar{f}(x,t) = 0 \tag{66}$$

which is the p.d.e. of the variations of the adjoint system.

X. EXTREMAL OF FUNCTIONAL FOR A SIMPLE OSCILLATOR. To show that $\delta^2 J$ must be positive semi-definite we start with an example by a simple harmonic oscillator with no forcing function. Thus from Eq. (63) we have the ordinary differential equation [6] .

$$\eta_{tt} + \omega^2\eta = 0 \tag{67}$$

The solution for the above equation is

$$\delta y = \mu\eta = A\cos(\omega t+\theta) \tag{68a}$$

$$\delta y_t = \mu\eta_t = -\omega A\sin(\omega t+\theta) \tag{68b}$$

Both A and θ can be determined from the following given initial conditions

$$\delta y(t=0) = \delta y(0) = A\cos\theta \tag{69a}$$

$$\delta y_t(t=0) = \delta yt(0) = -\omega A\sin\theta \tag{69b}$$

For computation by the finite element method the increment time is taken as T which gives

$$T = t_b - t_o = \left(\frac{n}{\omega}\right)\left(\frac{\pi}{2}\right) \tag{70}$$

where n = 1,2,3...

The image function becomes

$$\hat{\delta y}(t=T-t) = A\cos[\theta+\omega(T-t)] \tag{71a}$$

$$\hat{\delta y}_t(t=T-t) = -\omega A\sin[\theta+\omega(T-t)] \tag{71b}$$

For the ordinary differential equation we have the second variation from Eq. (52) which gives

$$\delta^2 J = \int_o^T [\hat{\delta y}_t(x,t=t)\hat{\delta y}_t(x,t=T-t)$$

$$+ \omega^2\hat{\delta y}(x,t=t)\hat{\delta y}(x,t=T-t)]dt \tag{72}$$

Separating Eq. (72) into two parts and using Eqs. (68) and (71) we have

$$\delta^2 J = \delta^2 J[\delta y_t] + \delta^2 J[\omega\delta y] \tag{73a}$$

where

$$\delta^2 J[\delta y_t] = \int_o^{\omega T} \omega^2 A^2 \sin(\theta+\omega t)\sin(\theta+\omega T-\omega t)d(\omega t) \tag{73b}$$

$$\delta^2 J[\omega\delta_y] = \int_o^{\omega T} \omega^2 A^2 \cos(\theta+\omega t)\cos(\theta+\omega T-\omega t)d(\omega t) \tag{73c}$$

and

$$\omega T = n(\frac{\pi}{2}) \tag{73d}$$

which is a multiple of $\pi/2$.

The trigonometric relationship for Eq. (73) is

$$\sin(\omega t+\partial) = \sin \omega t \cos \theta + \cos \omega t \sin \theta \tag{74a}$$

$$\cos(\omega t+\theta) = \cos \omega t \cos \theta - \sin \omega t \sin \theta \tag{74b}$$

$$\sin(\theta+\omega T-\omega t) = -\sin[\omega t - (\theta+n\pi/2)]$$

$$= -\sin \omega t \cos(\theta+n\pi/2) + \cos \omega t \sin(\theta+n\pi/2) \tag{74c}$$

and

$$\cos(\theta+\omega T-\omega t) = \cos[\omega t - (\theta+n\pi/2)]$$

$$= \cos \omega t \cos(\theta+n\pi/2) + \sin \omega t \sin(\theta+n\pi/2) \tag{74d}$$

For the case when n is odd, we have

$$\cos(\theta+n\pi/2) = (-1)^{\frac{n+1}{2}} \sin \theta \tag{75a}$$

$$\sin(\theta+n\pi/2) = (-1)^{\frac{n-1}{2}} \cos \theta \tag{75b}$$

For the case when n is even, we have

$$\cos(\theta+n\pi/2) = (-1)^{n/2} \cos \theta \tag{76a}$$

$$\sin(\theta+n\pi/2) = (-1)^{n/2} \sin \theta \tag{76b}$$

First, we take the case when n is odd. Substituting Eqs. (74) and (75) into Eq. (73), one obtains

$$\delta^2 J[\delta y_t] = \omega^2 A^2 \int_o^{n\pi/2} \{(\sin \omega t \cos \theta + \cos \omega t \sin \theta)$$

$$\cdot [-\sin \omega t(-1)^{(n+1)/2} \sin \theta + \cos \omega t(-1)^{(n-1)/2} \cos \theta]\}d(\omega t)$$

$$= (-1)^{(n-1)/2} \omega^2 A^2 \int_o^{n\pi/2} [\sin \theta \cos \theta + \sin \omega t \cos \omega t]d(\omega t)$$

$$= (-1)^{(n-1)/2} \omega^2 A^2 [\frac{1}{2} + \frac{n\pi}{2} \sin \theta \cos \theta] \tag{77a}$$

and

$$\delta^2 J[\omega\delta y] = \omega^2 A^2 \int_0^{n\pi/2} \{(\cos\omega t \cos\theta - \sin\omega t \sin\theta)$$

$$\cdot [\cos\omega t(-1)^{(n+1)/2} \sin\theta + \sin\omega t\, (-1)^{(n-1)/2} \cos\theta]\} d(\omega t)$$

$$= (-1)^{(n-1)/2} \omega^2 A^2 \int_0^{n\pi/2} [-\sin\theta\cos\theta + \sin\omega t \cos\omega t] d(\omega t)$$

$$= (-1)^{(n-1)/2} \omega^2 A^2 \left[\frac{1}{2} - \frac{n\pi}{2}\right) \sin\theta\cos\theta] \qquad (77b)$$

From Eq. (73a) when n is odd we have

$$\delta^2 J = (-1)^{(n-1)/2} \omega^2 A^2 \{[\frac{1}{2} + \frac{n\pi}{2} \sin\theta\cos\theta] + [\frac{1}{2} - \frac{n\pi}{2} \sin\theta\cos\theta]\}$$

$$\delta^2 J = (-1)^{(n-1)/2} \omega^2 A^2 \qquad (77c)$$

In particular for n = 1, one obtains

$$\delta^2 J = \omega^2 A^2 > 0 \qquad (78a)$$

which gives a minimum for the functional J. For n = 3

$$\delta^2 J = -\omega^2 A^2 < 0 \qquad (78b)$$

which gives a maximum for the functional J. It is noted that $\delta^2 J$ is independent of θ which is related to the starting conditions. It is also independent of the polarity of A since it appears in terms of A^2.

Now we take the case when n is even. Substituting Eqs. (74) and (76) into Eq. (73), one obtains

$$\delta^2 J[\delta y_t] = \omega\delta A^2 \int_0^{n\pi/2} \{(\sin\omega t \cos\theta + \cos\omega t \sin\theta)$$

$$[-\sin\omega t\, (-1)^{n/2} \cos\theta + \cos\omega t(-1)^{n/2} \sin\theta]\} d(\omega t)$$

$$= (-1)^{n/2} \omega^2 A^2 \int_0^{n\pi/2} [-\sin^2\omega t \cos^2\theta + \cos^2\omega t \sin^2\theta] d(\omega t)$$

$$= (-1)^{n/2} \omega^2 A^2 (-\cos^2\theta + \sin^2\theta) n\pi/4 \qquad (79a)$$

and

$$\delta^2 J[\omega\delta y] = \omega^2 A^2 \int_o^{n\pi/2} \{(\cos \omega t \cos \theta - \sin \omega t \sin \theta)$$

$$\cdot [\cos \omega t(-1)^{n/2} \cos \theta + \sin \omega t(-1)^{n/2} \sin \theta]\} d(\omega t)$$

$$= (-1)^{n/2} \omega^2 A^2 \int_o^{n\pi/2} [\cos^2 \omega t \cos^2 \theta - \sin^2 \omega t \sin^2 \theta] d(\omega t)$$

$$= (-1)^{n/2} \omega^2 A^2 (\cos^2 \theta - \sin^2 \theta) n\pi/4 \tag{79b}$$

From Eq. (73a) when n is even we have

$$\delta^2 J = (-1)^{n/2} \omega^2 A^2 \{(-\cos^2 \theta + \sin^2 \theta) + (\cos^2 \theta - \sin^2 \theta)\} n\pi/4$$

$$\delta^2 J = 0 \quad \text{for all } n = \text{even} \tag{79c}$$

We can conclude here that the functional J definitely [6] has a minimum if $\omega T = \pi/2$, or T is a quarter of the natural period of the oscillation $\tau = 2\pi/\omega$. Moreover, from Eq. (70) for n = 1

$$T = t_b - t_o = \pi/(2\omega) = \tau/4 \tag{80a}$$

If n = 2 and $\delta^2 J = 0$ in Eq. (79c), we have

$$T = t_b - t_o < \pi/\omega = \tau/2 \tag{80b}$$

This is the upper limit of the increment we chose for T, above which the minimum of the functional J is not guaranteed.

XI. EXTREMAL FOR A SIMPLY SUPPORTED BEAM WITH CONCENTRATED LOAD AT THE MIDDLE. To show that $\delta^2 J$ must be positive semi-definite we use the example of a simply-supported beam with a concentrated load at the middle. If the load is suddenly removed [7], Eq. (63) becomes

$$\eta_{tt} + b^2 \eta_{xxxx} = 0 \tag{81a}$$

Or from Eq. (56) we have

$$y_{tt} + b^2 y_{xxxx} = 0 \tag{81b}$$

The starting conditions are

$$\frac{d}{dt} \sigma_o(x) = 0 \tag{82a}$$

$$\sigma_o(x) = Mc/I = \frac{(Px/2)h}{I} \quad \text{for } 0 < x < \ell/2 \tag{82b}$$

and

$$\sigma_o(x) = \frac{P(\ell/2)h}{I} (1 - \frac{x}{\ell}) \quad \text{for } \ell/2 < x < \ell \tag{82c}$$

The solution for Eq. (81b) is

$$\sigma(x,t) = - Eh \frac{\partial^2 y}{\partial x^2}$$

$$= \frac{\delta\sigma_s}{\pi^2} \sum_{n=odd}^{\infty} \frac{1}{n^2} (-1)^{(n-1)/2} \sin(n\pi x/\ell)\cos p_n t \qquad (83a)$$

where

$$p_n = bn^2\pi^2/\ell^2 \qquad (83b)$$

and

$$\sigma_s = (P\ell/2I)(h/2) \qquad (83c)$$

The quantity σ_s is the initial static stress at the middle of the beam where $x = \ell/2$ and on the outer surface of the beam.

In order to find y_t we let

$$y = \frac{1}{Eh} \frac{8\sigma_s}{\pi^2} \sum_{n=odd}^{\infty} (\frac{1}{n^2})(\frac{\ell}{n\pi})^2(-1)^{\frac{n-1}{2}} \sin(n\pi x/\ell)\cos p_n t \qquad (84a)$$

Then by partial differentiation we have

$$\frac{\partial^2 y}{\partial x^2} = - \frac{1}{Eh} \frac{8\sigma_s}{\pi^2} \sum_{n=odd}^{\infty} (\frac{1}{n^2})(-1)^{\frac{n-1}{2}} \sin(n\pi x/\ell)\cos p_n t \qquad (84b)$$

which agrees with Eq. (83a), and

$$\frac{\partial y}{\partial t} = - \frac{1}{Eh} \frac{8\sigma_s}{\pi^2} \sum_{n=odd}^{\infty} (\frac{1}{n^2})(-1)^{\frac{n-1}{2}} b \sin(n\pi x/\ell) \sin p_n t \qquad (84c)$$

where from Eq. (83b)

$$b = p_n\ell^2/(n^2\pi^2) \qquad (84d)$$

and

$$p_1 = b\pi^2/\ell^2 \qquad (84e)$$

It is noted that the index n appears in both spatial and temporal functions in Eqs. (84a) and (84b) under the summation sign. We are interested in finding those functions of t that are independent of the index n. Let us assume that

$$p_n t = \frac{n^2\pi}{\ell} (\ell/2-c)$$

$$= n^2\pi/2 - n^2\pi c/\ell \qquad (85)$$

It is noted that for $n = 1, 3$, and 5, $n^2\pi/2$ becomes $\pi/2$, $4\pi + \pi/2$, and $12\pi + \pi/2$, respectively.

Thus we have

$$\begin{aligned} \cos p_n t &= \cos\,[n^2\pi/2 - n^2\pi c/\ell] \\ &= \cos\,[\pi/2 - n^2\pi c/\ell] \\ &= \sin\,(n^2\pi/c/\ell) \end{aligned} \tag{86a}$$

and

$$\sin p_n t = \cos\,(n^2\pi c/\ell) \tag{86b}$$

Moreover, for $c/(\ell/2) = 1$, $1/2$, and 0

$$\cos p_n t = \sin[(\pi/2)c/(\ell/2)] = 1,\ 0.707,\ \text{and}\ 0,\ \text{respectively} \tag{87a}$$

and

$$\sin p_n t = \cos[(\pi/2)c/(\ell/2)] = 0,\ 0.707,\ \text{and}\ 1,\ \text{respectively} \tag{87b}$$

The above functions are independent of index n at those values of $c/(\ell/2)$. Thus, Eq. (84) may be rewritten at those values as

$$\frac{\partial^2 y}{\partial x^2}(x,t=\hat{t}) = -\frac{\sigma_s}{Eh}\sin(\pi c/\ell)y_o(x) \tag{88a}$$

and

$$\frac{\partial y}{\partial t}(x,t=\hat{t}) = -\frac{\sigma_s}{Eh}\cos(\pi c/\ell)y_o(x) \tag{88b}$$

where

$$y_o(x) = \sum_{n=\text{odd}}^{\infty} (-1)^{\frac{n-1}{2}} \sin(n\pi x/\ell) \tag{88c}$$

The series terms in Eq. (88c) are the result of an expansion of a triangular deflection of the form

$$y_o(x) = x/(\ell/2) \qquad \text{for } 0 < x < \ell/2 \tag{88d}$$

$$y_o(x) = 2 - x/(\ell/2) \qquad \text{for } \ell/2 < x < \ell \tag{88e}$$

as shown in Figure 2.

For the images of Eqs. (84b) and (84c) the time dependent terms become

$$\cos p_n t(T-t) = \cos(p_n T-p_n t) \tag{89a}$$

and

$$\sin p_n(T-t) = \sin(p_n T-p_n t) \tag{89b}$$

The term p_nT can be obtained from Eqs. (84d) and (84e) as

$$p_nT = (bn^2\pi^2/\ell^2)T = n^2p_1T \tag{89c}$$

For computation by the finite element method the increment in time is taken as T which is defined as

$$T \overset{\Delta}{=} t_b - t_o = (m/p_1)(\pi/2) \tag{90a}$$

where

$$m = 1,2,3... \tag{90b}$$

Then with the aid of Eqs. (85), (89c), and (90a) we have

$$p_n(T-t) = mn^2(\pi/2) - [n^2\pi/2 - n^2\pi c/\ell]$$

$$= (m-1)n^2(\pi/2) + n^2c/\ell \tag{90c}$$

Then for the case when m = 1, the time dependent terms become

$$\cos p_n(T-t) = \cos(n^2\pi c/\ell) \tag{91a}$$

and

$$\sin p_n(T-t) = \sin(n^2\pi c/\ell) \tag{91b}$$

By similar method we can obtain

$$\frac{\partial^2 y}{\partial x^2}(x,\hat{t}=T-t) = -\frac{\sigma_s}{Eh}\cos(\pi c/\ell)y_o(x) \tag{92a}$$

$$\frac{\partial y}{\partial t}(x,\hat{t}=T-t) = -\frac{b\sigma_s}{Eh}\sin(\pi c/\ell)y_o(x) \tag{92b}$$

For the partial differential equation we have the second variation from Eq. (52) which gives

$$\delta^2 J = \int_o^T [\delta y_t(x,\hat{t}=t)\delta y_t(x,\hat{t}=T-t)$$

$$+ b^2\delta y_{xx}(x,\hat{t}=t)\delta yxx(x,\hat{t}=T-t)]dt \tag{93}$$

Separating Eq. (93) into two parts and using Eqs. (88) and (92), we have

$$\delta^2 J = \delta^2 J[\delta y_t] + \delta^2 J[b\delta y_{xx}] \tag{94}$$

The first term on the right of Eq. (94) is

$$\delta^2 J[\delta y_t] \cong \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o{}^2(x)dx \int_o^{p_1T=\pi/2} \cos(\pi c/\ell)\sin(\pi c/\ell)d(p_1t)$$

$$\cong \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o{}^2(x)dx > 0 \tag{95a}$$

where

$$p_1t = \pi/2 - \pi c/\ell \qquad d(p_1t) = -(\pi/\ell)dc \tag{95b}$$

at

$$p_1t = \pi/2 \quad , \quad c/(\ell/2) = 0 \quad , \quad \pi c/\ell = 0 \tag{95c}$$

at

$$p_1t = 0 \quad , \quad c/(\ell/2) = 1 \quad , \quad \pi c/\ell = \pi/2 \tag{95d}$$

The second term of Eq. (94) is

$$\delta^2 J[b\delta y_{xx}] = \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o{}^2(x)dx \int_0^{\pi/2} \sin(\pi c/\ell) \cos(\pi c/\ell) d(\pi c/\ell)$$

$$= \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o{}^2(x)dx > 0 \tag{95e}$$

Thus by combining Eqs. (95a) and (95e), one obtains

$$\delta^2 J = \int_{x_o}^{x_2} 2(b\sigma_s/Eh)^2 y_o{}^2(x)dx > 0 \tag{96}$$

which gives a minimum for the functional J.

Now we take the case when m = 2, Then Eqs. (90c) and (89) become

$$p_n(T-t) = n^2\pi/2 + n^2\pi c/\ell \tag{97a}$$

$$\cos p_n(T-t) = \cos(n^2\pi/2+n^2\pi c/\ell) \quad , \quad n^2 = 1,9,25, \text{ etc.}$$

$$= -\sin(n^2\pi c/\ell) \tag{97b}$$

and

$$\sin p_n(T-t) = \sin(n^2\pi/2+n^2\pi c/\ell) \quad , \quad n^2 = 1,9,25,\text{etc.} \tag{97c}$$

$$= \cos(n^2\pi c/\ell)$$

Thus the image function becomes

$$\frac{\partial^2 y}{\partial x^2}(x,\hat{t}=T-t) = -\frac{\sigma_s}{Eh}(-\sin \pi c/\ell)y_o(x) \tag{98a}$$

$$\frac{\partial y}{\partial t}(x,\hat{t}=T-t) = -\frac{\sigma_s}{Eh}\cos(\pi c/\ell)y_o(x) \tag{98b}$$

By substituting Eqs. (88) and (98) into Eq. (94) we have

$$\delta^2 J \cong \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o^2(x)dx \int_o^{\pi/2} [\cos^2(\pi c/\ell) - \sin^2(\pi c/\ell)]d(\pi c/\ell)$$

$$\cong \int_{x_o}^{x_b} (b\sigma_s/Eh)^2 y_o^2(x)dx(\frac{1}{2})[\sin \pi - \sin 0) = 0 \qquad (99a)$$

We can conclude here that the functional J definitely [8] has a minimum if $p_1 T = \pi/2$, where T is a quarter of the natural period of the oscillation $\tau = 2\pi/p_1$. Moreover, from Eq. (90a) for m = 1, we have

$$T = t_b - t_o = \pi/(2p_1) = \tau/4 \qquad (99b)$$

If m = 2 and $\delta^2 J = 0$ in Eq. (99a), we can conclude that

$$T = t_b - t_o < \pi/p_1 = \tau/2 \qquad (99c)$$

This is the upper limit of the increment we choose for T, above which the minimum of the functional J is not guaranteed.

XII. CONCLUSIONS. The functional in bilinear form is symmetrical about the original variables and the adjoint variables. The Euler Lagrange equations for the original and the adjoint systems are derived using the fundamental lemma of the calculus of variations. By integrating the bilinear expression by parts, one can obtain the bilinear concomitant in terms of initial and boundary terms. The adjoint system can be arranged in a manner that it is a reflected mirror of the original system in time. Thus the initial conditions for the bilinear concomitant become zero.

Generalized boundary conditions using many types of "springs" relating the various spatial partial derivatives are defined to satisfy the boundaries of the concomitant. The higher partials in original variables and variations in the adjoint variables can be kept in low orders by these "springs". Algorithms are developed for use in the finite element method by taking the first variations of the functional. These algorithms are simplified because the adjoint system gives exactly the same solutions as that of the original system.

Sensitivity coefficient is found to be similar to the variation of the variable and obeys the same partial differential equation. The solution of the original p.d.e. is taken as the solution of the variations for two examples, a simple oscillator and a simply-supported beam with load at the middle. It is found that the second variation of the functional is positive semi-definite if the increment in time for the finite element method is less than half the natural period of the physic systems in both cases. This will guarantee a minimum for the functional and thus the method is truly workable if employed as algorithms for the finite element method.

REFERENCES

1. Stacey, Weston, M. Jr., Variational Methods in Nuclear Reactor Physics, Academic Press, 1974.

2. Rund, H., The Hamilton-Jacobi Theory of the Calculus of Variations, Robert E. Krieger Publishing Company, Huntington, NY, 1973.

3. Gelfand, I. M. and Fonier, S. V., Calculus of Variations, Prentice-Hall, 1963.

4. Sagan, Hans, Introduction to the Calculus of Variations, McGraw-Hill, 1969.

5. Tomovic, Rajko, Sensitivity Analysis of Dynamic Systems, McGraw-Hill, 1963.

6. Shen, C. N. and Wu, Julian J., "A New Variational Method for Initial Value Problems, Using Piecewise Hermite Polynomial Spline Functions, ARO Report 81-3, Proceedings of the 1981 Army Numerical Analysis and Computers Conference, 1981.

7. Jacobsen, Lydiks and Ayre, Robert S., Engineering Vibrations, McGraw-Hill, 1958.

8. Shen, C. N., "Variational Principle for Gun Dynamics With Adjoint Variable Formulation," Proceedings of the Third US Army Symposium on Gun Dynamics, Volume II, p. IV-108, May 1982.

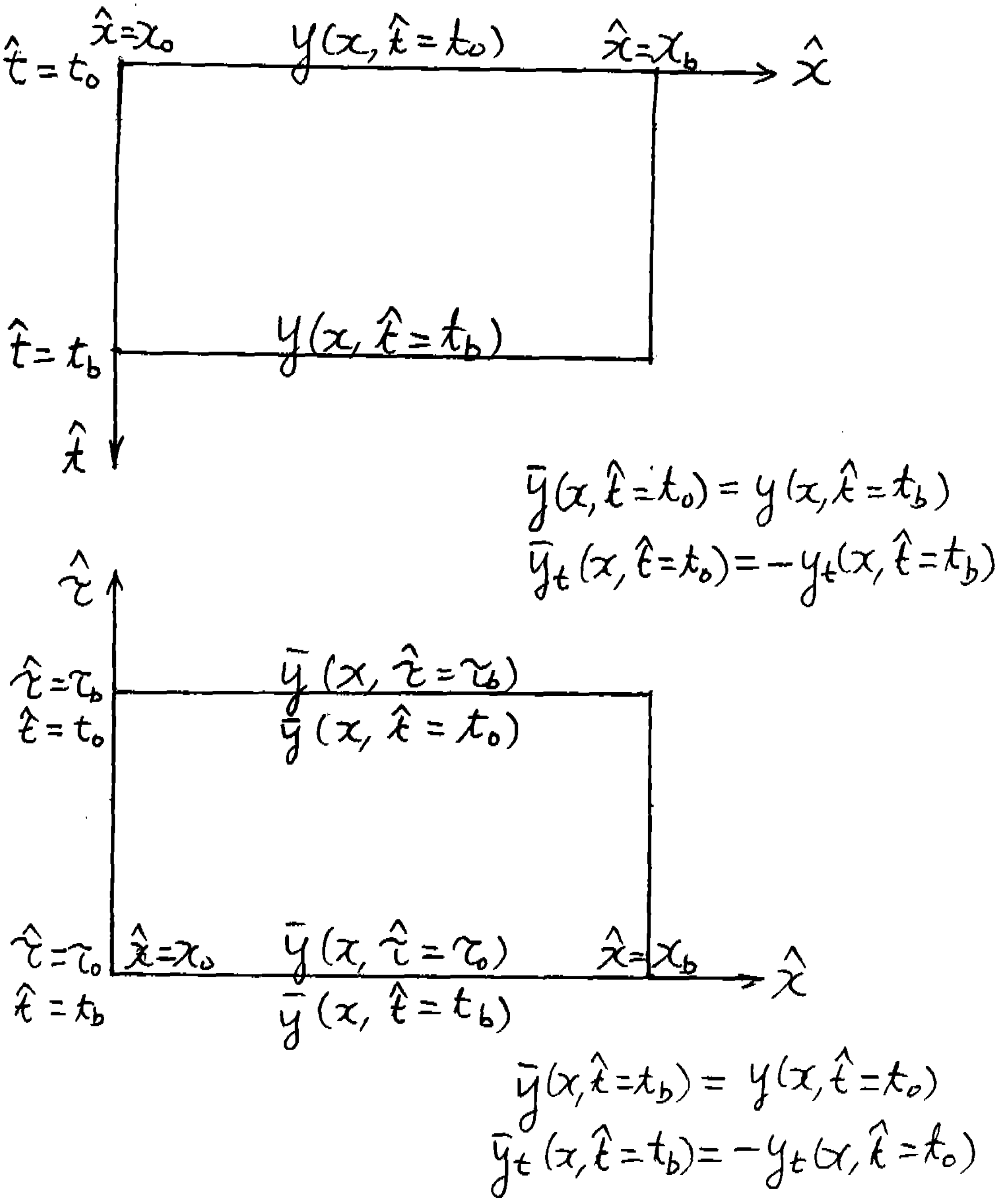


Figure 1. Image Reflection of the Adjoint System.

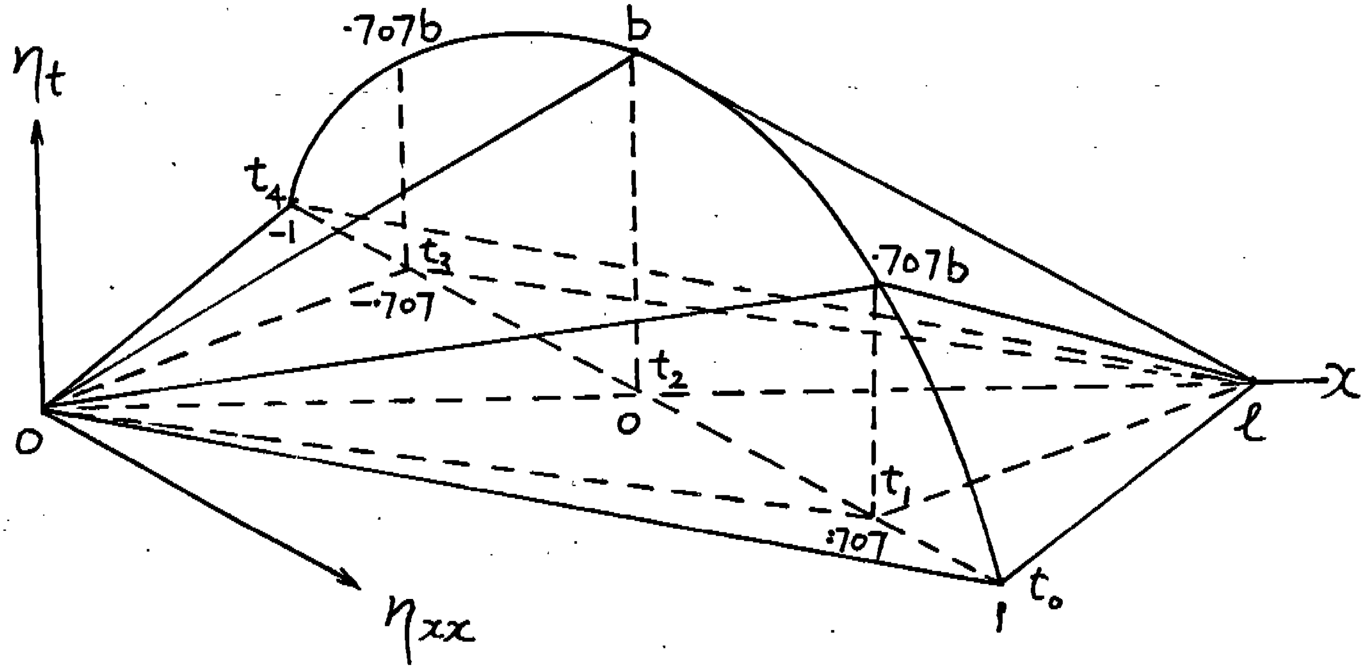


Figure 2. Variation of the Partials for a Beam Equation.

THE INFLUENCE OF CUTOUTS ON TRANSIENT RESPONSE OF HEMISPHERICAL SHELLS

AARON D. GUPTA
Mechanical Engineer

HENRY L. WISNIEWSKI
Mathematician
Ballistic Research Laboratory
U.S. Army Armament Research and Development Command
Aberdeen Proving Ground, MD 21005

ABSTRACT

To evaluate performance of enclosure structures due to an internal blast, the dynamic response of a continuous hemispherical configuration has been compared to the response of the same configuration, but with a large access opening. The results indicate substantial alteration of the response behavior and stress concentration effects in the vicinity of the cutout region.

INTRODUCTION

The Ballistic Research Laboratory is currently in the process of acquiring a target enclosure to facilitate destructive terminal ballistic testing of chemical explosives (CE), armor, and kinetic energy (KE) penetrators by safe containment of blast, fragments, and resultant harmful combustion products. The present investigation is an extension of the work reported in Reference 1, where a simplified continuous configuration was considered. A concept layout of the firing range is shown in Figure 1. A detailed description of the layout is given in Reference 2.

A salient feature of the structure is a large sliding door with a configuration to match the curvature of the hemispherical wall. The primary function of the door is to allow equipment access inside the enclosure. The door is sealed to the wall with a pressurized hose seal along its perimeter. Additionally, an air exhaust system mounted at the rear of the structure operates during the test and draws aerosolized material out of the enclosure after a test and traps it in filters in the exhaust ducting. The entire structure is built to contain blast and fragments, to trap aerosolized materials, and to permit photographic observation of the test.

For a continuous shell configuration clamped to a horizontal rigid foundation, the three-dimensional problem was reduced to a two-dimensional axisymmetric configuration by considering a single pie-shaped segment of the hemispherical enclosure. The entire structure was subsequently generated through 360-degree rotation of the segment about the vertical axis of symmetry resulting in quite economical computer calculations. However, the analysis suffered from a shortcoming inasmuch as the results could not be extrapolated to a discontinuous shell configuration with cutouts and access opening due to nonlinear geometric characteristics. The existence of high stress concentrations at the corners can reduce the margin of safety and result in permanent deformation in these regions.

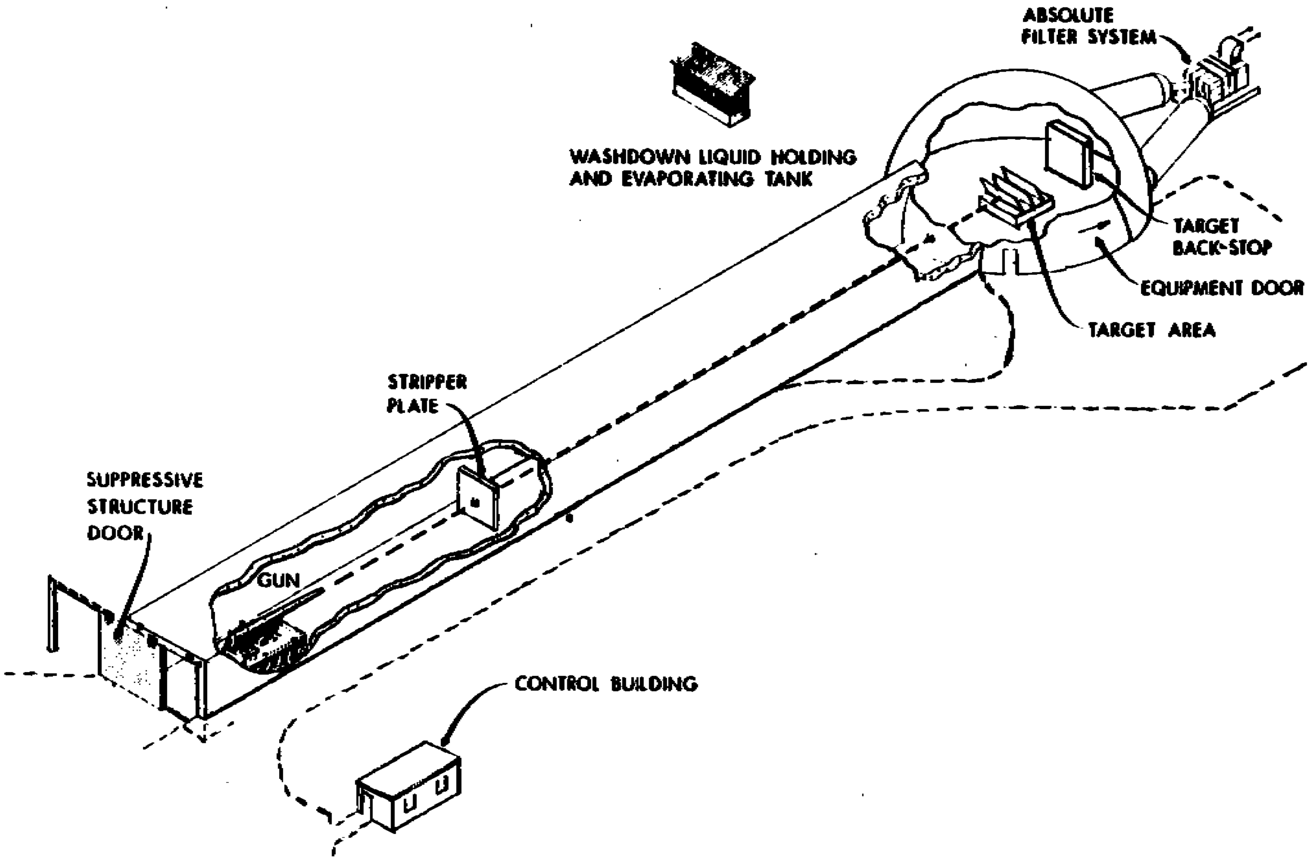


Figure 1. Preliminary concept layout of the AHKELS (Advanced High Kinetic Energy Launch System) range.

The structural integrity of a discontinuous configuration could not be assured with certainty from the study of a continuous structure and a detailed three-dimensional analysis with inclusion of the cutout was deemed to be necessary.

The current study is devoted to a demonstration of feasibility of modeling large scale structures with cutouts and determination of the influence of a relatively large cutout on the response behavior of a containment structure using a comparative evaluation of a discontinuous enclosure relative to a continuous configuration.

For an enclosure structure in a test facility, entrance hole opening for incoming projectiles and duct openings for filtration equipment at the rear of the structure are necessary in addition to personnel and equipment access openings. These openings are relatively small compared to the large rectangular equipment access door opening. Although these openings have been included to allow for venting of internal pressure to the ambient atmosphere, they are ignored in the simplified finite-difference model under the assumption that relative to the large door opening, they have rather small influence upon structural response. Taking advantage of this approximation and the resulting lateral symmetry, only one half of the structure and the door opening have been considered and multiple hole interaction effects have been excluded.

Previous work[3-5] on modeling of structures with cutouts involved relatively simple structures with symmetrically located cutouts of rather simple shapes. The current improvements in modeling techniques are significant insofar as large scale three dimensional structures with relatively large cutouts of various shapes, sizes, and locations subjected to complex loading conditions were analyzed without difficulty.

ESTIMATION OF TRANSIENT AND QUASI-STEADY LOADS

A major problem associated with the enclosed range tests is the overpressure that results[6] from the very rapid heating of air within the enclosure as the penetrator and the target are torn apart during the encounter. The structure and the seals must bear the load for the entire life cycle of the range without catastrophic failure of critical regions during loading and unloading phases until the pressure is vented and the products are cooled through mass and heat loss mechanisms to ambient conditions.

Since the key elements of the firing ranges are the enclosure structures, the structural analysis group of the Blast Dynamics Branch at the Ballistic Research Laboratory (BRL), was assigned the task of estimating the overpressure loading on the wall and assure structural integrity from a conservative viewpoint. The choice of a hemispherical configuration was due to an earlier investigation by N. J. Huffington, et.al.[7] who studied a continuous hemispherical configuration.

Since quasi-steady and dynamic pressures upon the wall are not expected to be significantly altered due to the closed equipment door during a test, the internal loading data in Reference 1 was used for both continuous and discontinuous structures. The salient features of the 25.4mm thick hemispherical shell design and the contoured equipment access door, 4.267m wide and 5.486m high are described in References 2 and 8.

DYNAMIC RESPONSE ANALYSIS

Response of both continuous and discontinuous structures subjected to internal pressurization from a centrally located explosive blast was simulated using a BRL version of the PETROS 3.5 computer program[9] which employs the finite-difference method to solve the nonlinear equations governing finite-amplitude elasto-plastic response of thin Kirchhoff shells. The model is valid for large deflections and can be employed to treat the entire structure rather than a small section.

Material Model. The uniaxial tensile quasi-static stress-strain property of 1020 steel described in Reference 7 was used for primary vessel material in this analysis. The material was modeled in the code as a combination of three linear segments followed by a perfectly-plastic behavior and linear elastic unloading, resulting in a polygonal approximation of the experimental data. The strain hardening part of the stress-strain curve is generated by a sublayer hardening model[9] from a weighted combination of elastic perfectly-plastic curves yielding a piecewise multilinear hardening representation. Strain-rate effects were neglected, which is conservative since these effects tend to increase the structural resistance and thus reduce the total deformation.

Finite Difference Model. To add capability to handle discontinuous geometry, the PETROS 3.5 code was modified substantially at BRL. For computational purposes, the boundary conditions from the outer edges of the shell configuration were applied selectively to the inner boundaries of the cutout. The boundary conditions at the edges of the cutout can be symmetric, clamped, hinged, or free or any combinations of these. Capability for modeling multiple cutouts is available to facilitate study of interaction effects provided a minimum number of meshpoints are allowed between cutouts.

To simplify the model of the enclosure with the rectangular cutout, only one half of the cutout and the structure was considered due to lateral symmetry. A total of 22 meshpoints in the circumferential and 20 meshpoints in the hoop direction were used, except in the cutout regions wherein only 12 meshpoints were used. Each meshpoint used for Gaussian integration points through a single thickness layer. The meshpoints at the base of the structure and along the periphery of the cutout were restrained from movement in both axial and transverse directions. Clamped edge conditions were imposed along the cutout boundary along which the door must be tightly clamped to the enclosure wall during a test. All other meshpoints were allowed unrestrained movement in any direction. Only the nodes were subjected to blast pressure in the radially outward direction.

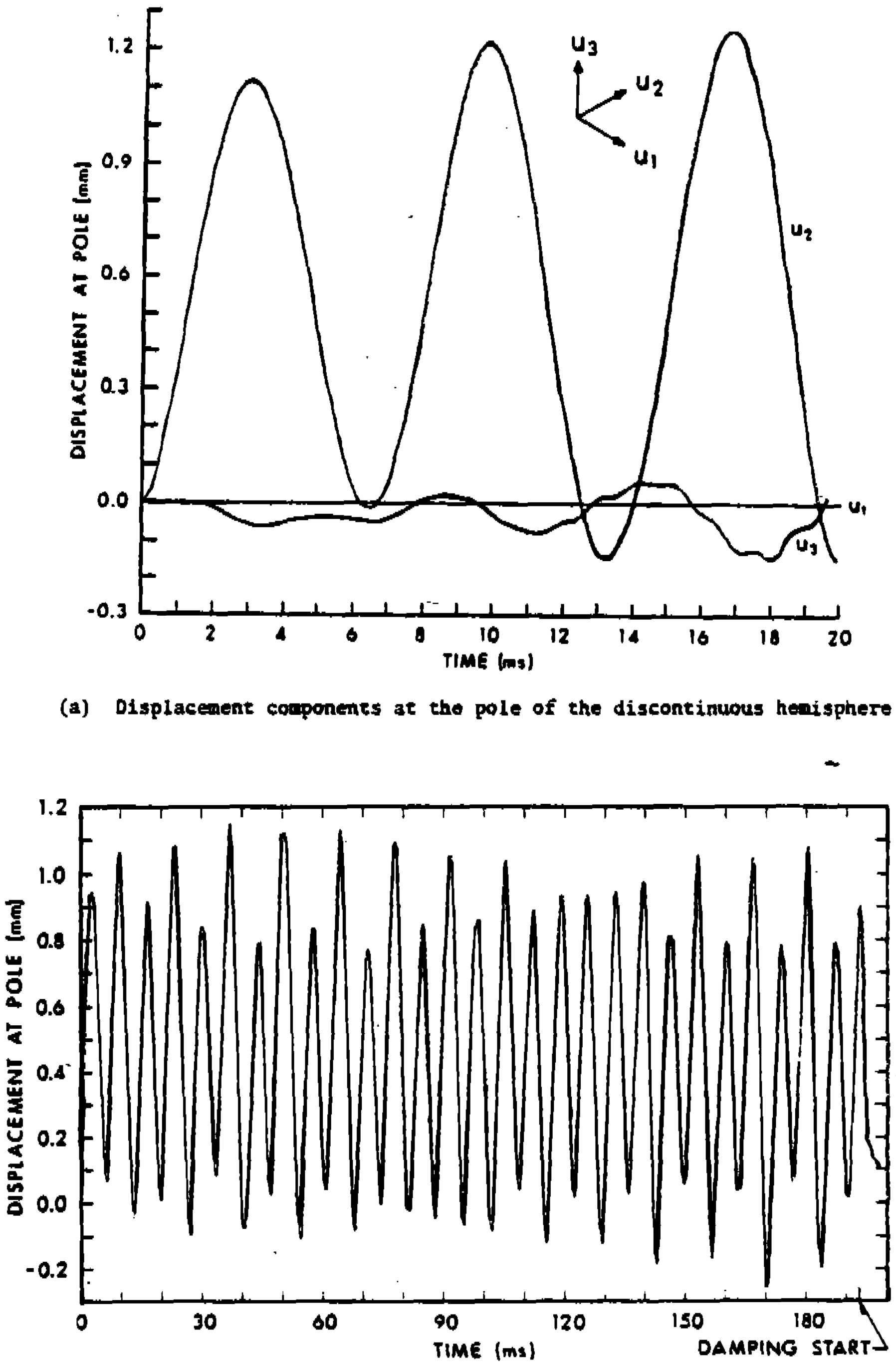
For the continuous structure, only a quarter segment was modeled using 18 equal width meshes along the surface and a single layer through the thickness to represent the pie shaped segment. Four Gaussian integration points through the thickness were used at each meshpoint.

RESULTS AND DISCUSSION

The deformed model of the discontinuous hemispherical shell configuration relative to the undeformed configuration shows a peak deflection of 1.28mm in the radially outward direction at approximately 17 ms in Figure 2a which is comparable to a peak displacement of 1.17mm observed at approximately 36 ms for the continuous enclosure shown in Figure 2b. However, in addition to the radially outward displacement, the discontinuous configuration exhibits a lateral displacement component with a highly reduced peak of .16mm indicating a tendency of the pole to shift laterally by a small amount. This behavior is in marked contrast to the single degree of freedom motion in the radial direction for the continuous configuration[1] where the pole exhibits no lateral movement. The resultant peak deflection, which is obtained by vectorial sumation of individual displacement components, is found to be 1.29mm predominantly in the radially outward direction at the pole for the discontinuous enclosure and is nearly 9% larger than corresponding polar deflection for the continuous structure with identical geometric and material parameters. However, this displacement is less than 4.5% of the shell thickness for both structures so that geometric nonlinearities are insignificant. The peak deflection is of the order of elastic deflection at the pole and the residual deformation at the pole is negligible after elastic oscillations are damped out and internal pressure is released.

Energy balance studies for both configurations using the code confirmed absence of plastic work and numerical instability. Both total and kinetic energies were bounded. The fluctuations of kinetic energy appeared to have twice the frequency of the work performed by the internal blast pressure.

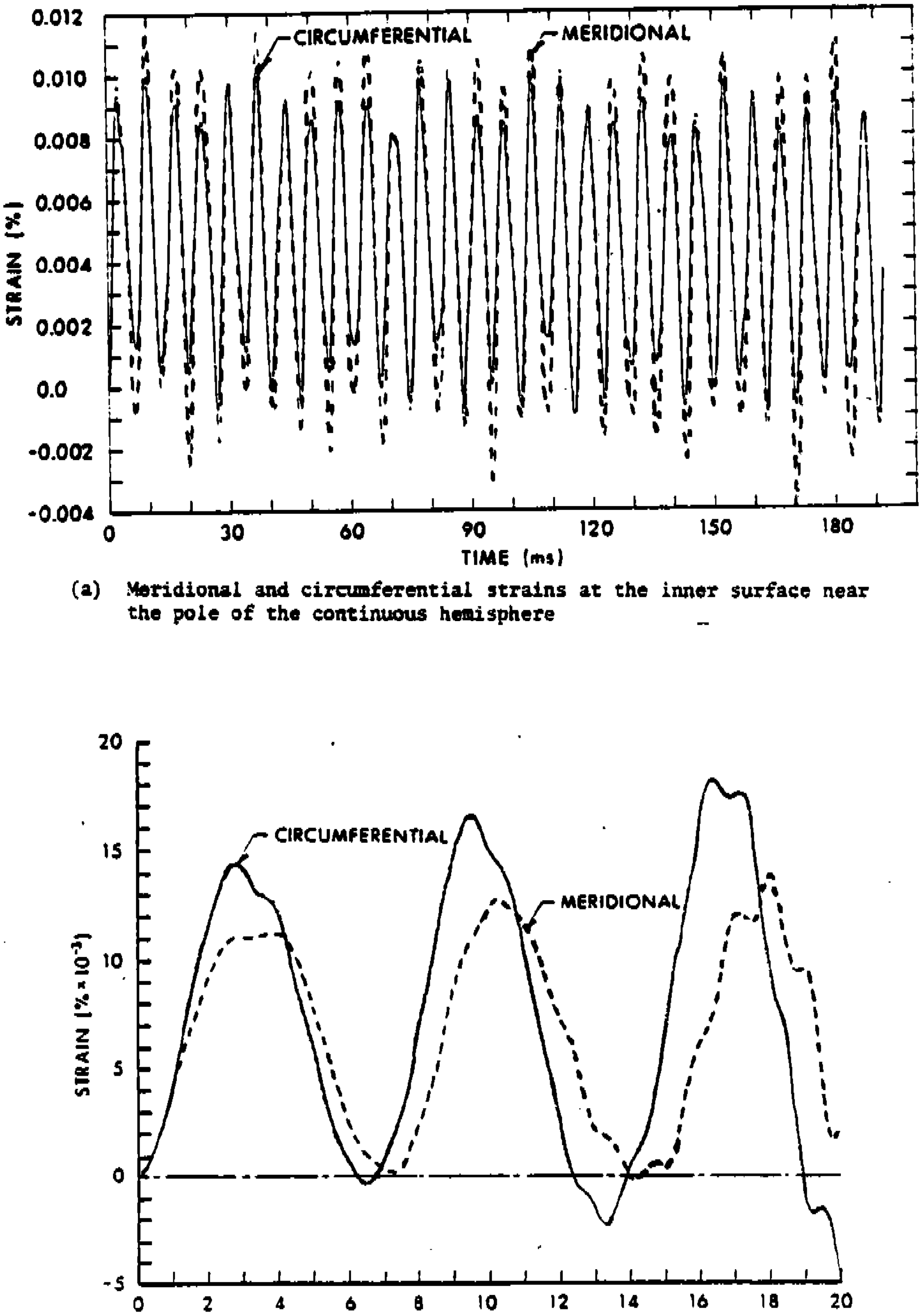
Transient circumferential and meridional surface strains at both outer and inner walls near the pole for the continuous hemispherical enclosure[1] are in phase and approximately equal in magnitude indicating domination by membrane effects due to elastic vibration of the wall in the breathing mode as shown in Figure 3a for the inner surface. Strain components at the clamped



(a) Displacement components at the pole of the discontinuous hemisphere

(b) Displacement at the pole of the continuous hemisphere

Figure 2. Transient displacements at the pole of the continuous and discontinuous hemispheres.



(a) Meridional and circumferential strains at the inner surface near the pole of the continuous hemisphere

(b) Meridional and circumferential strains at the inner surface near the pole of the discontinuous hemisphere

Figure 3. Comparison of surface strains near the pole of the continuous and discontinuous hemispheres.

edge at the base are in general out of phase and unequal in magnitude due to flexural deformation. Substantial weakening of the flexural wave was observed near the pole. However, for the discontinuous shell configuration, noticeable bending effect in addition to stretching was observed as far as one mesh away from the pole. Typical strain response at this location at the inner surface is shown in Figure 3b.

Strains were computed at the midpoint of the top edge of the rectangular cutout as well as at the corner point to assess the influence of the cutout on the response behavior of the structure. Meridional strains were approximately three times as high as the circumferential strains at both outer and inner surfaces at the midpoint of the top edge of the cutout. A peak meridional strain of .019% at approximately 10 ms was observed at the inner wall in contrast to .015% at the outer wall. A similar response due to combined stress concentration and bending effect from the clamped edges was observed at the corner. However, meridional strains were in general of the same order of magnitude as the circumferential strains at this location at both outer and inner walls.

At a point one mesh away from the pole, circumferential and meridional strains were nearly equal and in phase. A peak circumferential strain of .020% was observed at 18 ms at this location at the outer wall in contrast to a maximum of .018% strain at the inner wall indicating propagation of bending wave from the clamped cutout edge to the pole. Thus both peak strains and deflection were found to occur simultaneously at the pole for a discontinuous configuration. This is in direct contrast to the behavior of the doorless enclosure configuration in which peak strains developed at the clamped edge at the base due to combined bending and stretching while peak deflections occurred at the pole due to focusing of vibratory energy. The development to stress concentration and bending effects of the clamped cutout edges. However, in all cases strains were small enough to be in the linear elastic regime and peak stresses calculated from elastic theory did not exceed 40 MPa. Margins of safety at critical regions based on a yield strength of 241 MPa for mild steel were equal to or greater than 5.0 which was found to be satisfactory.

ACKNOWLEDGEMENTS

This investigation was performed for Mr. Louis Giglio-Tos, who was the Project Coordinator for Project No. T01400 at R-9. Valuable assistance from Mr. Frederick H. Gregory and Dr. Joseph M. Santiago of the Terminal Ballistics Division is gratefully acknowledged.

REFERENCES

1. Gupta, A. D. and Wisniewski, H. L., "Dynamic Response of a Hemispherical Enclosure Subjected to Explosive Blast Loading," ARO Report 82-1, May 1982, Transactions of the Twenty-Seventh Conference of Army Mathematicians, West Point, NY.

2. Gupta, A. D. and Wisniewski, H. L., "Dynamic Analysis of a Hemispherical Enclosure Subjected to Explosive Blast Loading," Computer Analysis of Large Scale Structures - Part II, AMD, Vol. 49, November 19, 1981, ASME Winter Annual Meeting, Washington, DC.

3. Paramasivan, P., "Free Vibration of Square Plates with Square Openings," Journal of Sound and Vibration, Vol. 30(2), 1973, pp. 173-178.

4. Aksu, G. and Ali, R., "Determination of Dynamic Characteristic of Rectangular Plates with Cutouts Using a Finite Difference Formulation," Journal of Sound and Vibration, Vol. 44(1), 1976, pp. 147-158.

5. Brogan, F., Forsberg, K., and Smith, S., "Dynamic Behavior of a Cylinder with a Cutout," AIAA Journal, Vol. 7, No. 5, May 1969, pp. 903-911.

6. Abrahams, R., Peterson, R., and Bertrand, B., "Measurement of Blast Pressure Produced by Impact of Kinetic Energy Penetrator on a Steel Target," BRL Memorandum Report ARBRL-MR-02983, January 1980, AD B045141L.

7. Huffington, N. J. and Robertson, S. R., "Containment Structure Versus Suppressive Structures," BRL Memorandum Report 2597, February 1976.

8. Hill, W. V., "Design of the Advanced High Kinetic Energy Launch System," Aberdeen Proving Ground, Proceedings of the 20th Explosives Safety Meeting, 23-27 August 1982, Norfolk, VA.

9. Pirotin, S. D., Berg, B. A., and Witmer, E. A., "PETROS 3.5: New Developments and Program Manual for the Finite-Difference Calculation of Large Elastic-Plastic Transient Deformations of Multi-Layer Variable-Thickness Shells," Ballistic Research Laboratories Contract Report 211, February 1975.

THE COUPLING OF TRACKED AND INTERIOR WAVES IN A FRONT TRACKING SCHEME

J. Glimm[1,2] O. McBryan[3,4] B. Plohr[4] S. Yaniv[2]

Courant Institute of Mathematical Sciences
New York University
New York, NY 10012

ABSTRACT. In a front tracking scheme certain waves are given a more exact treatment through the introduction of a dynamically moving, lower dimensional grid that is aligned with the wave fronts. These tracked waves are propagated in time through the solution of Rankine-Hugoniot relations; the remaining waves are propagated using finite difference equations. In this paper we discuss the coupling between these two wave systems.

1. INTRODUCTION.

Front tracking[3,6] is a method for numerical solution of hydrodynamical problems whose basic element is a system of moving curves (i.e. the front). Each curve represents a physical wave. For gas dynamics, these waves are either shock waves (nonlinear sound waves) or contact discontinuities (temperature jumps or slip lines). The motion of each type of wave is determined by Rankine-Hugoniot equations that relate the states on the two sides of the wave. In the interior regions, i.e. the portions of space disjoint from the curves, the solution is thought of as being smooth (or relatively smooth, with only small discontinuities). Therefore the interior solution can be calculated well using finite difference methods.

Two further issues remain to be specified before the front tracking scheme is defined. The first, which is the subject of this paper, is the interaction between tracked and interior waves. The second, not discussed here, is the interaction between two or more intersecting tracked waves, where the tracked waves do not define a problem that is locally one-dimensional.

2. THE NATURE OF THE COUPLING OF INTERIOR AND TRACKED WAVES.

During a time step $[t,t + \Delta t]$, interior waves may reach one of the tracked waves. When this occurs, the interior wave can be transmitted, reflected, and absorbed. Also the tracked wave can spontaneously radiate waves into the interior. (For example, curvature of the front causes this.) Within the tracked wave, there is both normal and tangential motion. The normal motion is specified by the Rankine-Huguniot equations. The tangential motion corresponds to surface waves and curvature effects.

[1] Supported in part by the National Science Foundation, Grant MCS-8207965.
[2] Supported in part by the Army Research Office, Contract No. DAAG29-83-K-0007.
[3] Alfred P. Sloan Foundation Fellow.
[4] Supported in part by the Department of Energy, Grant DEA 027ER03077.

For computational simplicity, we wish to decouple the front and interior calculations, as much as it is scientifically correct to do so. The extent of the front-interior coupling is determined by the domain of dependence of the moving front. Stated differently, we need to know all waves within a distance

$$x = O(\Delta t)$$

of the front in order to propagate the front through the time interval Δt. To accomplish this, we introduce local normal and tangential coordinates, and then use a fractional step (operator splitting) method to advance the front, first normally and then tangentially.

3. THE DATA STRUCTURE. The data structure of the computation consists of state variables (density, etc.) specified at points of a regular rectangular grid, together with double-valued state variables (left density, right density, etc.) specified at irregular mesh points on a moving front. To define the state variables at an arbitrary point $\vec{x}$ we interpolate between these regular and irregular points. Because the front divides the plane into connected components, it is important to interpolate only data corresponding to the component to which $\vec{x}$ belongs; the following procedure is used.

First the regular mesh block containing $\vec{x}$ is determined. If all four corners have the same component as $\vec{x}$ then the state at $\vec{x}$ is given by bilinear interpolation. If three or fewer corners have the same component as $\vec{x}$ then the mesh block is triangulated, the triangle in which $\vec{x}$ is located is found, and the state at $\vec{x}$ is calculated by linear interpolation from the states at the corners of this triangle. The triangulation is chosen so that each triangle lies entirely within a single component.

4. THE NORMAL SWEEP TO PROPAGATE THE FRONT. At each mesh point on the front, a normal direction $\hat{n}$ is defined. The solution at time t is evaluated at this mesh point (where it is double valued, so that two states, left and right, are obtained). The solution at time t is also evaluated at a normal distance Δx on each side of the front. These states will be used to calculate the waves that impinge on the front from the interior. Notice that if the front contains curves too close to each other, these new evaluation points $\vec{x} \pm \Delta x * \hat{n}$ may be in different components from the mesh point $\vec{x} \pm 0 * \hat{n}$ at the front. In this case the evaluation point $\vec{x} \pm \Delta x * \hat{n}$ is shifted into the correct component. (Conceptually, the state at a point outside a given component is obtained by extrapolation.) Thus we always have as data two left states corresponding to one component and two right states corresponding to a second component.

These states are used as initial data for an extended or non-local Riemann problem. Higher order solutions of these Riemann problems have been discussed before. To reduce sampling error in the random choice method, a steady state Ansatz has been used[4,2] to extend local Riemann data over a mesh block, thereby obtaining a non-local Riemann problem, which is solved to higher order. Higher order Godunov schemes[1] also employ ideas related to solutions of Riemann problems. In the present scheme we solved the non-local Riemann problem as follows.

Using the left and right states located at the front we solve an ordinary Riemann problem. The solution is the correct answer to the non-local problem at time t + 0, so it is used to approximate the propagation speeds of the characteristics that enter the two sides of the tracked wave. By following these characteristics backward from t + Δt to t we find their starting points in the normal intervals

$$[\vec{x} \pm 0 * \hat{n}, \vec{x} \pm |\Delta x| * \hat{n}].$$

The states at these points are calculated using linear interpolation between the states on the front and the states at the normally displaced points. In this way we determine which waves from the normal intervals enter the front. Using differential equations in characteristic form we compute corrected left and right states on the front at time t + Δt - 0. These corrected states define a new Riemann problem, whose solution gives states to be associated with the propagated front at time t + Δt.

5. THE TANGENTIAL SWEEP ALONG THE FRONT. By linear interpolation, the doubled valued variables can be defined at all the points of the front. Therefore we can evaluate the normally propagated solution at each mesh point and at two neighboring points displaced a distance Δx along the front. Using these three stencil points and the one-dimensional Lax-Wendroff scheme we determine the tangentially propagated state variables. Notice that tangential propagation of the points on the front is equivalent to a remeshing of the front, in the limit Δx -> 0, so it is not essential to move these points during the tangential sweep.

6. THE INTERIOR SCHEME NEAR TRACKED WAVES. For the calculation of the solution in the interior regions we use the two-dimensional Lax-Wendroff scheme, which involves two half steps. To facilitate the coupling of the front and interior, the front is also advanced in half steps, so its position and state variables are known at t + $\Delta t/2$.

Since the Lax-Wendroff scheme is a leapfrog composition of two Lax-Friedrichs steps it is enough to describe the Lax-Friedrichs scheme. This scheme usually assumes the initial data to be known at the four corners of a square. In our application, however, there are irregular squares, i.e. ones for which one or more of its corners is cut off by the front and thus lies in the wrong component. To circumvent this difficulty we view the Lax-Friedrichs scheme as defined by a flux balance: the sum of the fluxes through the sides of a mesh block determines the change in time of a conserved quantity integrated over the block. From this viewpoint the Lax-Friedrichs scheme is defined for irregular squares as well as regular squares. In fact, the propagation of the front also determines the fluxes through the front and thereby through the sides of the irregular squares.

7. TESTS AND VALIDATION. The present implementation of the above scheme is incomplete, with the coupling of the interior scheme to the tracked waves (Sec. 5) replaced by a more primitive version. Full implementation and validation represents work in progress. Here we show preliminary results for the supersonic flow past a wedge obstacle in a channel, using for comparison the steady-state solution obtained using the method of characteristics[5]. In this flow a bowshock interacts with a Prandtl-Meyer expansion.

The test flow was simulated using a 50 by 50 grid for 200 time steps. (The stability requirements for the finite difference schemes dictate that the downstream signal speed be approximately .5 grid blocks per time step.) The initial data was a slight perturbation of the steady state solution obtained using the method of characteristics. In Figs. 1a and 1b we show the initial and final shock positions and isopycnic (constant density) contours. In Figs. 2a and 2b we show the initial and final density distributions along two sides of the shock. Figs. 3a,b and 4a,b show the analogous distributions along the wedge and along the portion of the exit below the shock, respectively. These figures indicate that the front tracking scheme accurately reproduces the steady state solution.

1. P. Colella and P. Woodward, "The Piecewise Parabolic Method (PPM) for Gas-Dynamical Simulation," J. Comp. Phys., To appear (1983).

2. H. Glaz and T.-P. Liu, "The Asymptotic Analysis of Wave Interactions and Numerical Calculations of Transonic Nozzle Flow," Preprint (1982).

3. J. Glimm, E. Isaacson, D. Marchesin, and O. McBryan, "Front Tracking for Hyperbolic Systems," Adv. Appl. Math., Vol. 2, p. 91 (1981).

4. J. Glimm, G. Marshall, and B. Plohr, "A Generalized Riemann Problem for Quasi-One-Dimensional Gas Flows," Adv. Applied Math., To appear (1983).

5. G. Marshall and B. Plohr, "The Random Choice Method for Two-Dimensional Steady Supersonic Shock Wave Diffraction Problems," Preprint, New York University (1983).

6. B. Plohr, J. Glimm, and O.McBryan, "Applications of Front Tracking to Two-Dimensional Gas Dynamics Calculations," in Lecture Notes in Engineering Vol. 3, ed. J. Chandra and J. Flaherty, Springer-Verlag, New York (1983).

Fig. 1a: isopycnic contours at step 0

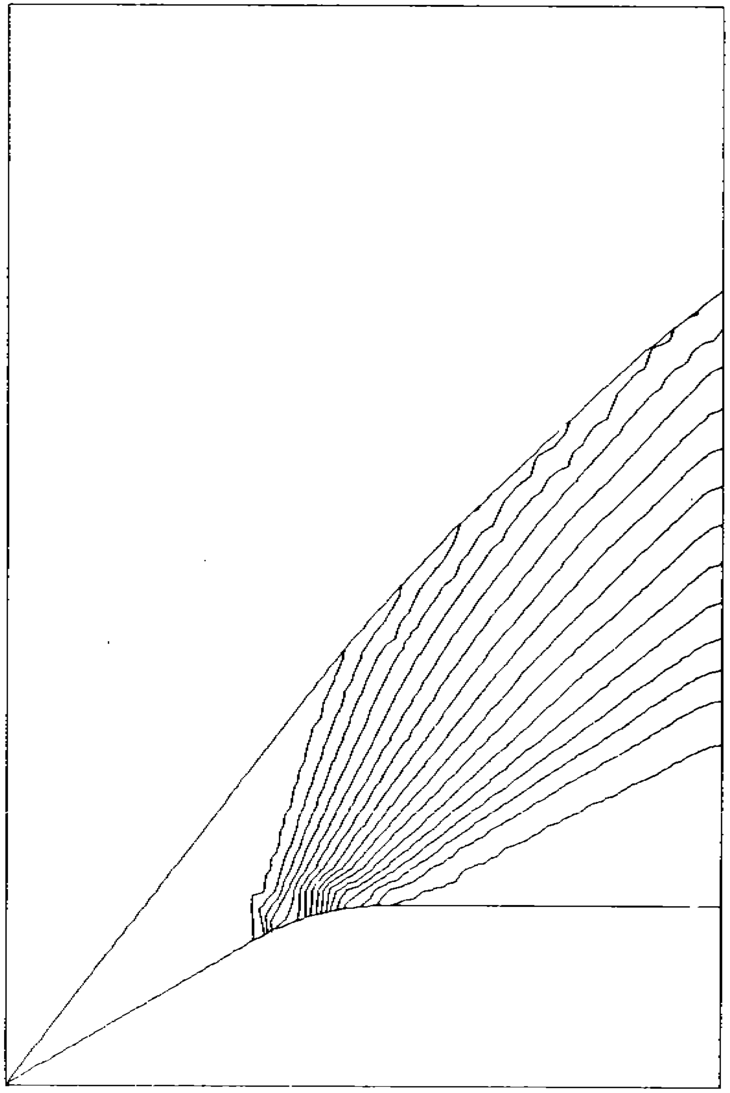
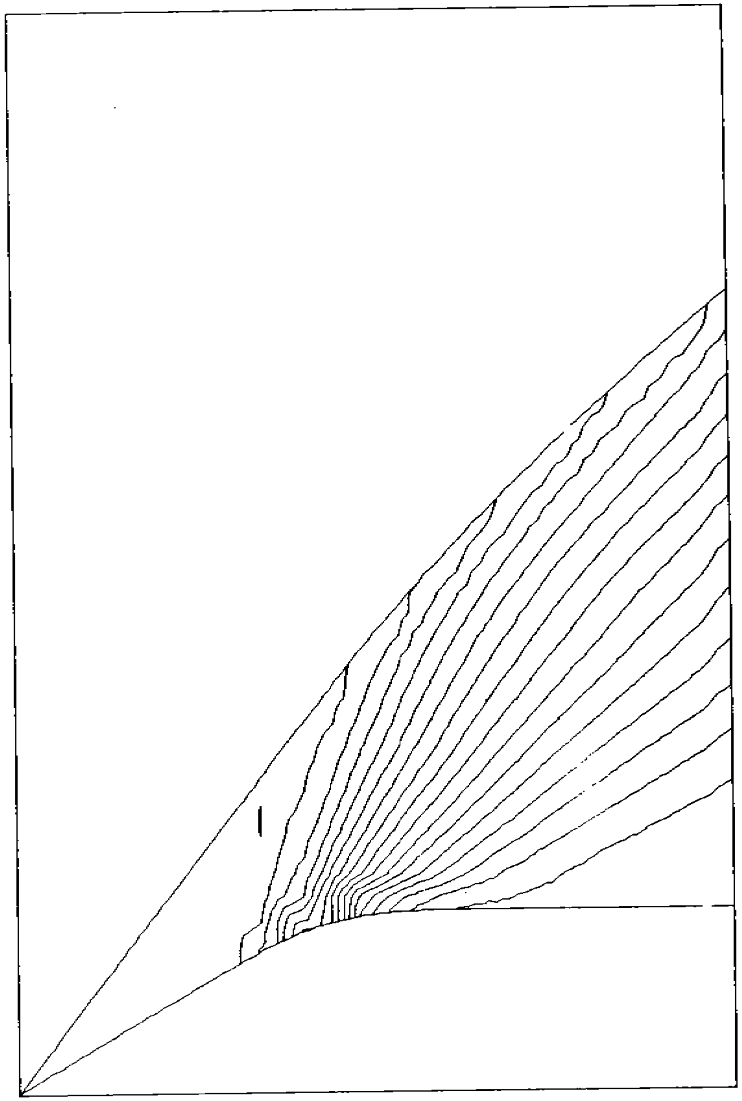
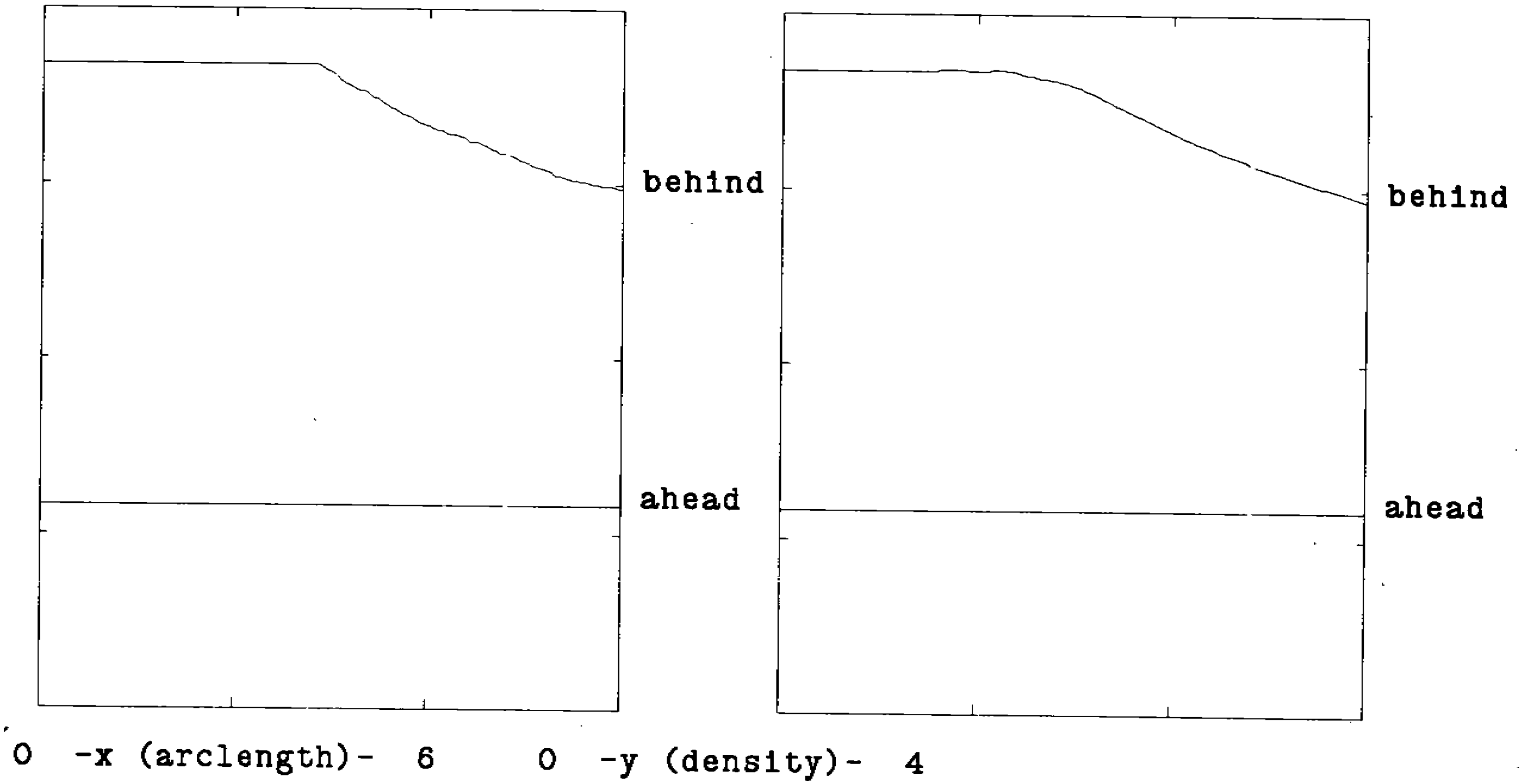


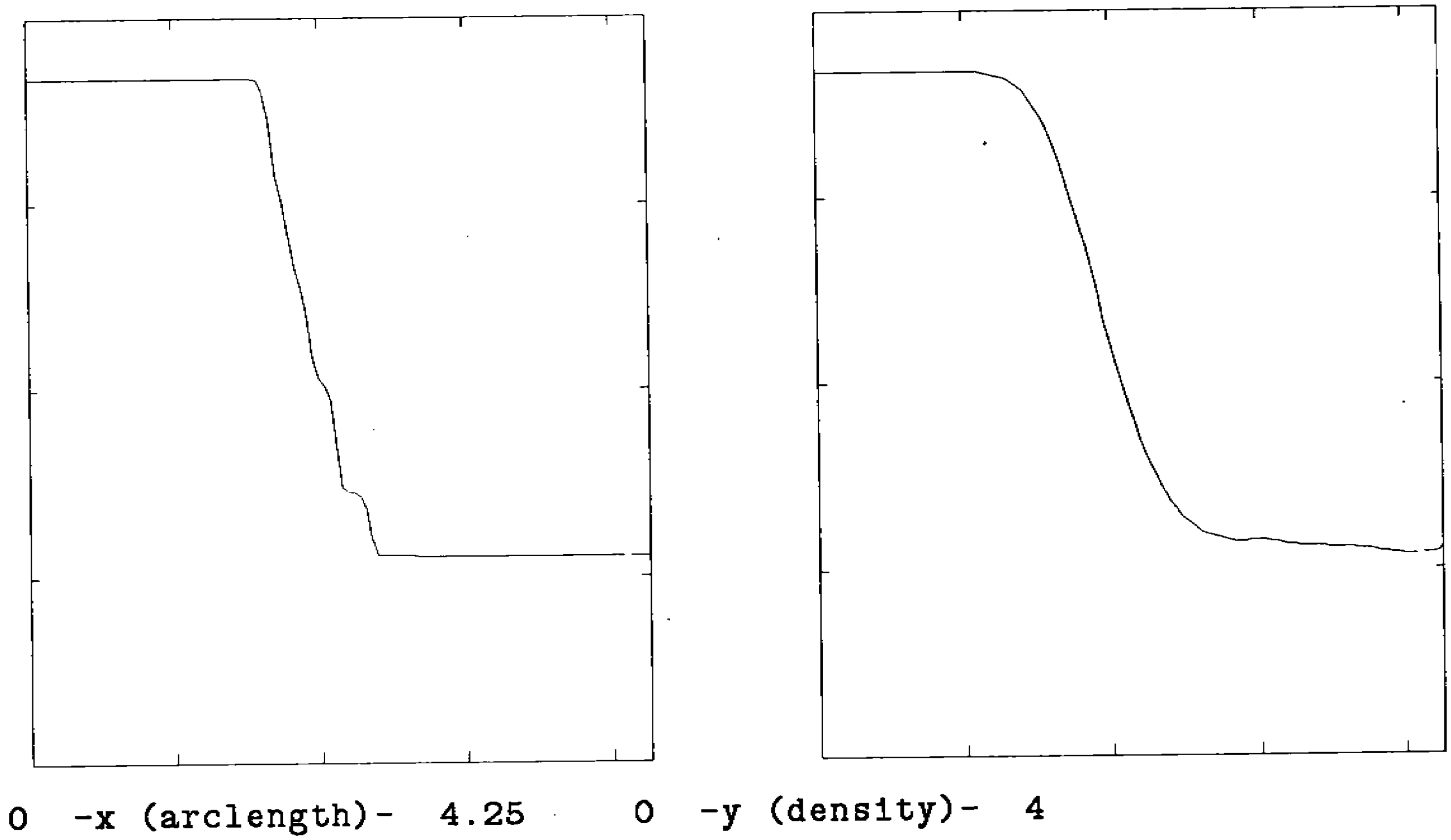
Fig. 1b: isopycnic contours at step 200



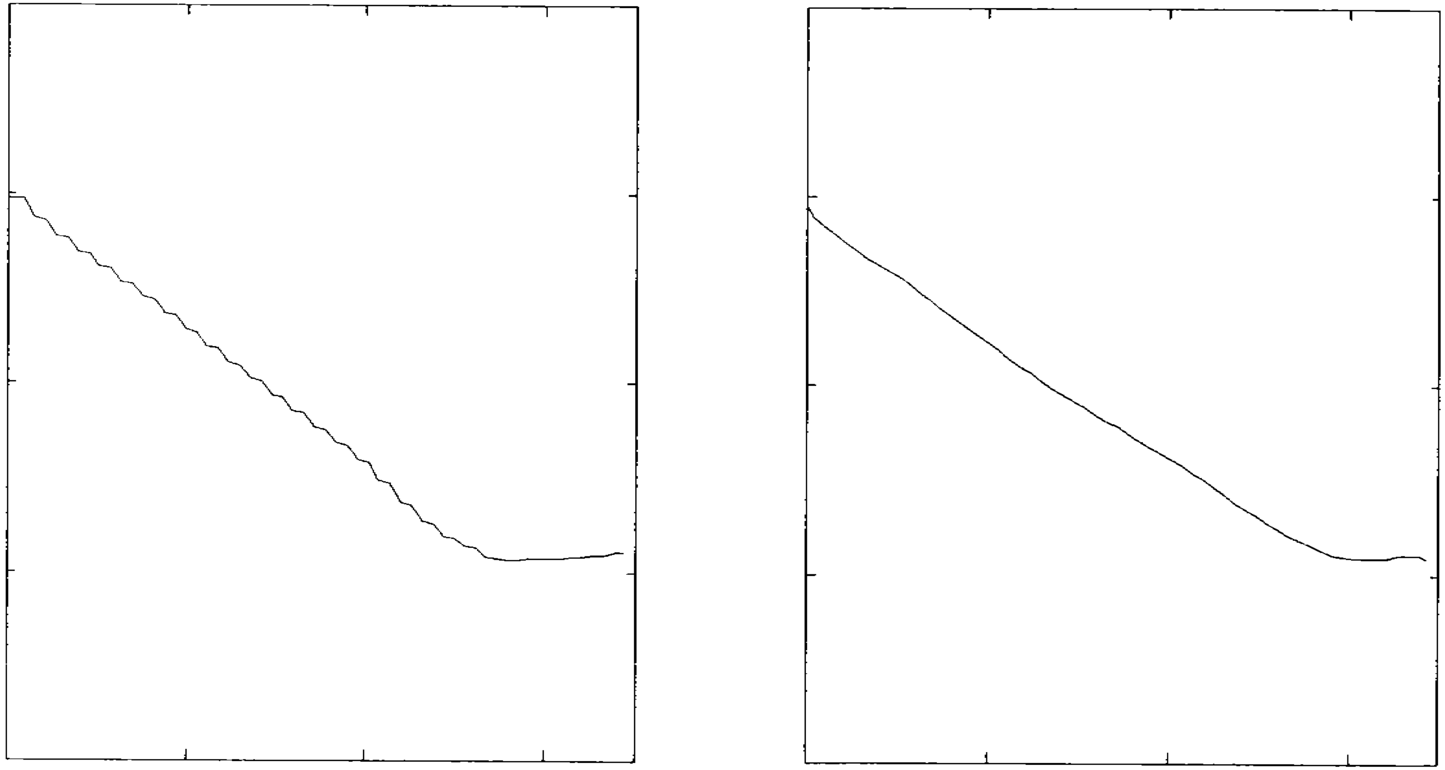
Figs. 2a,b: density along shock at steps 0 and 200



Figs. 3a,b: density along wedge at steps 0 and 200



Figs. 4a,b: density along exit at steps 0 and 200



0 -x (arclength)- 3.5   0 -y (density)- 4

RIEMANN SOLVERS, THE ENTROPY CONDITION AND HIGH RESOLUTION DIFFERENCE APPROXIMATIONS

Stanley Osher
Department of Mathematics
University of California
405 Hilgard Avenue
Los Angeles, CA 90024

Research Supported by: ARO Grant DAAG29-82-0090

A condition on the numerical flux for approximations to scalar, nonconvex conservation laws is introduced, and shown to guarantee convergence to the correct physical solution. These considerations lead to a simple, closed form, analytic expression for the solution to the Riemann problem for scalar, nonconvex, conservation laws.

A systematic approach is presented for converting these first order accurate convergent approximations to second order accurate, variation diminishing, entropy condition, satisfying approximations. The technique is extended to systems of equations of inviscid, compressible flow in general geometries, using a high resolution version of the author's scheme. The results of such multidimensional calculations are given in [2].

The work described in the second paragraph is joint with S. Chakravarthy [2].

For a single scalar conservation law

$$(1) \qquad u_t + f(u)_x = 0 \ , \quad 0 < t < \infty \ , \quad -\infty < x < \infty$$

for general nonconvex $f(u)$ we consider the Riemann problem.

$$\begin{aligned} u(x,0) &\equiv u^L \ , \quad x < 0 \\ &\equiv u^R \ , \quad x > 0 \end{aligned}$$

u^L, u^R arbitrary constants.

It is well known e.g. [3] that there exists a unique solution to (1) of the form

$$u(x,t) = u\left(\frac{x}{t}\right) = u(\zeta)$$

taking on values between u^L and u^R and satisfying the entropy condition.

We present here a closed form expression for the solution via:

<u>Lemma 1</u>. If $u^L < u^R$, then

$$(2) \qquad (a) \quad f(u(\zeta)) - \zeta u(\zeta) = \min_{w \in [u^L, u^R]} [f(w) - \zeta w]$$

if $u^L > u^R$,

then

$$(b) \quad f(u(\zeta)) - \zeta u(\zeta) = \max_{w \in [u^R, u^L]} [f(w) - \zeta w] \ .$$

Theorem 1.

(3) (a) If $u^L < u^R$, then

$$u(\zeta) = -\frac{d}{d\zeta}\left(\min_{w\in[u^L,u^R]} [f(w) - \zeta w]\right)$$

(b) If $u^L > u^R$, then

$$u(\zeta) = -\frac{d}{d\zeta}\left(\max_{w\in[u^R,u^L]} [f(w) - \zeta w]\right).$$

For proofs, see [2]. In that reference, we also obtain a simple condition on the numerical flux for semi-discrete approximations to (1), which guarantees convergence to the correct physical solution. This seems to be the most general class of convergent schemes known.

BIBLIOGRAPHY

[1] S. OSHER, Riemann Solvers, the entropy condition, and difference approximations. SINUM (to appear).

[2] S. OSHER, and S. CHAKRAVARTHY, High resolution schemes and the entropy condition, (submitted to SINUM)

[3] P.D. LAX, Shock waves and entropy, in "Contributions to nonlinear functional analysis", E.H. Zarontonello, Editor, Acad. Press, New York (1971), pp. 603-634.

ARTIFICIAL MASS CONCEPT AND TRANSONIC VISCOUS FLOW EQUATION

George S. Dulikravich
Department of Aerospace Engineering
and Engineering Mechanics
University of Texas at Austin
Austin, Texas 78712

Peter Niederdrenk
Institut Für Theoretische
Strömungsmechanik
DFVLR-AVA
Göttingen, F.R. Germany

ABSTRACT. By varying the grid clustering on the surface of an airfoil, it was observed that symmetric shocked solutions develop with a nonunique shock strength and location when numerically solving the full potential equation.

It is shown analytically that the conventional form of artificial density (or viscosity) produces a number of truly nonlinear terms which are suspected to be the cause of the nonuniqueness for all the finite grid sizes. A concept of artificial mass flow is shown to be suitable for analytically evaluating a new exact form of the switching function that eliminates all the nonlinear terms for any value of the local Mach number. The resulting expanded full potential equation then becomes a third order partial differential equation of permanently parabolic type resembling Sichel's transonic viscous flow equation. Consequently, our expanded full potential equation does not require the introduction of the customarily used artificial time concept.

I. INTRODUCTION. The numerical techniques for solving full potential equation modelling transonic flows are presently based on two similar concepts: an explicitly added artificial viscosity [1] and an implicitly modified artificial density [2]. Both techniques should create additional terms in the full potential equation in such a way that they nullify the numerical error introduced when using upstream rotated finite differencing in the locally supersonic regions.

If these artificial dissipative terms are introduced in a divergence free form, it was believed that the numerical solution will be unique [3]. Nevertheless, in recent years it was observed that the question of uniqueness is not resolved when isentropic discontinuities (shocks) are present in the solution. It has been shown that the finite difference formulation [4] of the artificial viscosity can somewhat affect the location and the strength of the isentropic discontinuities without causing numerical instability problems. The artificial damping was monotonically introduced in conservative form across the sonic line at all subsonic parts of the flow domain where the local value of the Mach number exceeded an arbitrary prescribed cut-off value [5]. A disturbing conclusion was that the numerical solution depends upon the choice of the initial guess for the potential field, varied choices for which causing the numerical solution to converge to strikingly different answers [5] if shocks are present in the field. At the same time it has been demonstrated that the nonunique solutions are not dependent on the type of the grid used nor are they dependent on the grid resolution for the grid sizes commonly used. The conclusion was that these were the correct nonunique solutions of the exact full potential equation.

We numerically experimented with a number of two-dimensional full potential cascade codes [6,7,8] and consistently observed non-unique symmetric shocked solutions. The test case was a nonstaggered cascade of NACA0012 airfoils spaced at 3.6 chord lengths apart. The free stream Mach number was $M = 0.8$, and the free stream angle of attack was zero. First order artificial viscosity in a fully conservative form [3,9] was used in each of the codes. Numerical solutions were obtained on a sequence of four successively refined O-type and C-type non-orthogonal boundary fitting geometrically periodic grids. When the grid points on the airfoils surface were equidistantly spaced, the iterative procedure converged to a symmetric solution with the shock located at approximately 80 percent of the chord.

Then the same sequence of grids with the same number of grid points was used with the only exception that the grid points on the airfoil's surface were symmetrically clustered closer to the leading edge and the trailing edge. The result was a symmetric solution with the shocks located at approximately 72 percent of the chord. It should be noted that the boundary and the periodicity conditions were enforced explicitly and exactly and that the airfoil surface Mach number drop across the shocks was in both test cases in close agreement with the exact one-dimensional isentropic shock conditions. These results were obtained with both O-type and C-type grids while using the same (symmetric) initial guess for the potential field, the same number of iterations and the same relaxation factors on each of the grids.

II. <u>ANALYSIS OF ARTIFICIAL DENSITY CONCEPT</u>. The following derivations should suggest the probable cause for nonuniqueness of the shocked solution of the discretized form of the full potential equation.

Mass conservation equation

$$\vec{\nabla}\cdot(\rho\vec{\nabla}\phi) = (\frac{\partial}{\partial s}\hat{e}_s + \frac{\partial}{\partial n}\hat{e}_n)\cdot(\rho\phi_{,s}\hat{e}_s + \rho\phi_{,n}\hat{e}_n) = 0 \qquad (1)$$

where comma denotes partial differentiation, can be written in its canonical form [10]

$$\vec{\nabla}\cdot(\rho\vec{\nabla}\phi) = \rho[(1-M^2)\phi_{,ss} + \phi_{,nn}] = 0 \tag{2}$$

If all the flow variables are nondimensionalized with their respective critical properties then

$$\frac{\rho_{,s}}{\rho} = -\frac{M^2}{M_*}M_{*,s} \tag{3}$$

If an artificial density of the general form

$$\tilde{\rho} = \rho - \Delta s\ \mu\rho_{,s} \tag{4}$$

is introduced, it can be shown that

$$\frac{\tilde{\rho}_{,s}}{\tilde{\rho}} = \frac{\rho_{,s}}{\rho} + \frac{1}{1 + \Delta s\ \mu\frac{M^2}{M_*}M_{*,s}}\left[\Delta s\ \mu\frac{M^2}{M_*}M_{*,s} + \Delta s(\mu\frac{M^2}{M_*})_{,s}M_{*,s}\right] \tag{5}$$

Hence, the mass conservation equation becomes

$$\vec{\nabla}\cdot(\tilde{\rho}\vec{\nabla}\phi) = \tilde{\rho}\{[1-M^2)\phi_{,ss} + \phi_{,nn}] + \Delta s\ \mu M^2\ \phi_{,sss} + E\} = 0 \tag{6}$$

where

$$E = \frac{\Delta s\ M_{*,s}}{1 + \Delta s\ \mu\frac{M^2}{M_*}\phi_{,ss}}\left[(\mu\frac{M^2}{M_*})_{,s} - (\mu\frac{M^2}{M_*})^2\Delta s\ M_{*,ss}\right] \tag{7}$$

Then the undesirable term in the expanded full potential equation is

$$E = \frac{\Delta s\ M^2}{1 + \Delta s\ \mu \frac{M^2}{M_*} \phi_{,ss}} [-\Delta s\ \mu^2 \frac{M^2}{M_*} \phi_{,ss}\ \phi_{,sss}$$

$$+ \mu((1+\gamma)\frac{M^2}{M_*^2} - 1)\frac{(\phi_{,ss})^2}{M_*} + \mu_{,s}\ \phi_{,ss}] \qquad (8)$$

The switching function μ is customarily [1,2] assigned the value

$$\mu = \mu_{JH} = 1 - \frac{1}{M^2} \qquad (9)$$

which is obtained from the condition that the term

$$\Delta s\ \mu\ M^2\ \phi_{,sss} \qquad (10)$$

should be approximately the same magnitude as the terms introduced by the upstream differencing of $\phi_{,ss}$ in the locally supersonic regions of the flow field. If μ_{JH} is used in E, the result is a cluster of truly nonlinear terms

$$E = \frac{\Delta s\ M^2}{1 + \Delta s(M^2-1)\ \frac{\phi_{,ss}}{\phi_{,s}}} [-\Delta s(M^2-1)(1- \frac{1}{M^2})\ \frac{\phi_{,ss}\ \phi_{,sss}}{\phi_{,s}}$$

$$+ (1- \frac{1}{M^2})((1+\gamma)\ \frac{M^2}{(\phi_{,s})^2} - 1)\ \frac{(\phi_{,ss})^2}{\phi_{,s}} + (1+\gamma)\ \frac{(\phi_{,ss})^2}{(\phi_{,s})^3}]$$

$$(11)$$

It is obvious that the conventional value of μ fails to make $E = 0$ for any value of local flow Mach number equal to or greater than one. Actually, when using μ_{JH} one ends up solving a nonlinear partial differential equation even on the sonic line where eq. 6 reduces to

$$\phi_{,nn} + \Delta s(1+\gamma)(\phi_{,ss})^2 = 0 \tag{12}$$

The substantiated explanation of the influence of the nonlinear terms on the final solution of the expanded full potential equation is not available at the present time. Nevertheless, it could be speculated that these nonlinear terms have some properties of solitons thus causing the numerically observed non-unique solutions for all realistic non-zero grid sizes. The influence of the nonlinear terms can be certainly affected by the particular finite differencing applied in the evaluation of the derivatives constituting the artificial viscosity [4] or artificial density [11]. Different shocked solutions can also be achieved by varying the expression for the switching function μ [12,13].

Full potential equation does not involve any dissipative mechanism on the basis of which expansion discontinuities could be eliminated. The numerically created artificial viscosity [1] and density [2] terms coupled with an artificial time concept [1] represented the basis of almost all transonic potential flow computations performed over the past decade.

By allowing for an insignificant vorticity generation in a limiting process applied to a small disturbance transonic potential equation, Sichel [14] has derived a "viscous transonic equation."

$$K_v \phi_{,xxx} + (K_\infty - \phi_{,x})\phi_{,xx} + \phi_{,yy} = 0 \tag{13}$$

where

$$K_v = (1 + \frac{\gamma-1}{Pr})/[\tau(1+\gamma)M_\infty^2]^{2/3}Re \tag{14}$$

and

$$K_\infty = (1 - M_\infty^2)/[\tau(1+\gamma)M_\infty^2]^{2/3} \tag{15}$$

Here, M_∞ is the free stream Mach number, γ is the ratio of specific heats, Pr and Re are Prandtl and Reynolds numbers, respectively, and τ is half the airfoil thickness ratio.

Since $K_v > 0$ this equation is parabolic and Chin [15,16,17] solved it numerically without any need to introduce the artificial time. Although it is better suited for modelling the physical details at the shocks, eq. 13 cannot be recast in a divergence free form.

III. ARTIFICIAL MASS FLUX CONCEPT. It is possible, nevertheless, to derive an expanded full potential equation that will always be of a parabolic type and will be readily expressible in a divergence free form. The idea is to expand the mass flux (rather than density or the speed of sound [2] alone) in a Taylor series. Such a modified mass conservation equation

$$\vec{\nabla}\cdot(\overline{\rho\vec{\nabla}\phi}) = 0 \tag{16}$$

can be expressed in the locally streamline aligned orthogonal coordinate system (s,n) as

$$\left(\frac{\partial}{\partial_s}\hat{e}_s + \frac{\partial}{\partial_n}\hat{e}_n\right)\cdot[(\rho\phi_{,s} - \Delta s\ \mu(\rho\phi_{,s})_{,s})\hat{e}_s + (\rho\phi_{,n})\hat{e}_n] = 0 \tag{17}$$

Hence,

$$\rho[(1-M^2)\phi_{,ss} + \phi_{,nn}] - \Delta s\ \rho[\mu_{,s}(\phi_{,ss} + \frac{\rho_{,s}}{\rho}\phi_{,s})$$

$$+ \mu(\frac{\rho_{,ss}}{\rho}\phi_{,s} + 2\frac{\rho_{,s}}{\rho}\phi_{,ss} + \phi_{,sss})] = 0 \tag{18}$$

Because

$$\rho = [\frac{\gamma+1}{2} - \frac{\gamma-1}{2}(\phi_{,s})^2]^{1/(\gamma-1)}$$

$$\tag{19}$$

$$\phi_{,s} = M_*$$

It follows that the artificial mass produces

$$\rho\{[(1-M^2)\phi_{,ss} + \phi_{,nn}] + \Delta s\ \mu(M^2-1)\phi_{,sss}\}$$

$$- \Delta s\ \rho\{(1-M^2)\mu_{,s}\phi_{,ss} - \mu[(1+\gamma)\frac{M^4}{M_*^3} + \frac{M^2}{M_*} - \frac{M^4}{M_*}](\phi_{,ss})^2\} = 0 \tag{20}$$

The second brace contains undesirable nonlinear terms. The value of the switching function μ is determined in such a way as to eliminate them entirely. Hence,

$$\frac{du}{\mu} = \frac{-2M_*}{(M_*^2-1)}\ \frac{d\ M_*}{(\frac{\gamma+1}{2} - \frac{\gamma-1}{2}M_*^2)} - (\frac{2}{\gamma+1})\ \frac{M_*\ dM_*}{(M_*^2-1)}$$

$$+ (\frac{2}{\gamma+1})\ \frac{M_*^3}{(M_*^2-1)}\ \frac{d\ M_*}{(\frac{\gamma+1}{2} - \frac{\gamma-1}{2}M_*^2)} \tag{21}$$

The result is

$$\mu = \frac{1}{(M_*^2-1)}\ (a^2)^{\frac{2-\gamma}{\gamma-1}} \tag{22}$$

Because of the relation

$$\frac{M^2-1}{M_*^2-1} = (\frac{\gamma+1}{2})(a^2)^{-1} \tag{23}$$

it follows that the modified mass conservation becomes

$$\rho\{[(1-M^2)\phi_{,ss} + \phi_{,nn}] + \Delta s(\frac{\gamma+1}{2})(\frac{1}{\rho})\phi_{,sss}\} = 0 \tag{24}$$

This expanded full potential equation is of parabolic type for any value of Mach number and its variable diffusion coefficient does not vanish for any finite value of the Mach number. Note the striking similarity between eq. 24 and eq. 13 and the fact that eq. 24 can be integrated without a need for artificial time variable [1]. The divergence free form of eq. 24 can be achieved as follows.
Let

$$A = \rho\phi_{,s} - \Delta s\ \mu(\rho\phi_{,s})_{,s} \tag{25}$$

Hence,

$$A = \phi_{,s}[\rho - \Delta s\ \mu(1 - \frac{1}{M^2})\rho_{,s}] \tag{26}$$

and the modified mass conservation (eq. 16) can be expressed in its fully conservative form as

$$\vec{\nabla}\cdot(\tilde{\rho}\vec{\nabla}\phi) = (\tilde{\rho}\phi_{,x})_{,x} + (\tilde{\rho}\phi_{,y})_{,y} = 0 \tag{27}$$

where the exact form of the artificial density is

$$\tilde{\rho} = \rho - \Delta s\ \mu(\frac{M^2-1}{M^2})\rho_{,s} \tag{28}$$

or

$$\tilde{\rho} = \rho - \Delta s(\frac{\gamma+1}{2})\ \frac{1}{M^2}\ (\frac{1}{\rho})\rho_{,s} \tag{29}$$

IV. <u>SUMMARY</u>. It has been analytically proven that the usual formulation of artificial density and viscosity terms leads to an introduction of truly nonlinear terms whose effects are suspected of causing certain numerical errors and inconsistencies in the numerical computation of transonic potential flows. A new concept of artificial mass flux was shown to produce only linear artificial dissipation that has the same basic character as a governing equation for physically viscous transonic flow. The artificial mass can be easily reformulated in terms of a new artificial density in a fully conservative form.

V. ACKNOWLEDGEMENTS. The Authors are thankful to the College of Engineering of the University of Texas at Austin for partially supporting this research. Special thanks are due to Professor Tinsley Oden of the Department of Aerospace Engineering and Engineering Mechanics for encouraging the continuation of the present work.

REFERENCES

1. Jameson, A., "Iterative Solution of Transonic Flows over Airfoils and Wings, Including Flows at Mach 1," Comm. Pure and Appl. Math, Vol. 27, 1974, pp. 283-309.

2. Hafez, M., South, J. and Murman, E., "Artificial Compressibility Methods for Numerical Solutions of Transonic Full Potential Equation," AIAA Journal, Vol. 17, No. 8, August 1979, ppl 838-844.

3. Jameson, A. and Caughey, D.A., "A Finite Volume Method for Transonic Potential Flow Calculations," Proceedings of AIAA 3rd Computational Fluid Dynamics Conference, Albuquerque, N. Mex., June 1977, pp. 35-54.

4. Chen. L.-T., and Caughey, D. A., "On Various Treatments of Potential Equations at Shocks," NASA CP 2201, 1981, pp. 121-138.

5. Steinhoff, J. and Jameson, A., "Multiple Solutions of the Transonic Potential Flow Equation." AIAA Journal, Vol. 20, No. 11, November 1982, pp. 1521-1525.

6. Dulikravich, D. S., "CAS2D-FORTRAN Program for Nonrotating Blade-to-Blade, Steady, Potential Transonic Cascade Flows," NASA TP 1705, 1980.

7. Dulikravich, D. S. and Sobieczky, H., "CAS22-FORTRAN Program for Fast Design and Analysis of Shock-Free Airfoil Cascades Using Fictitious-Gas Concept," NASA CR3507, 1982.

8. Dulikravich, D. S. and Sobieczky, H., "CAS24-FORTRAN Program for Shock-Free Design and Analysis of Transonic Realistic Cascades," to appear in 1983.

9. Dulikravich, D. S. and Sobieczky, H., "Shockless Design and Analysis of Transonic Cascade Shapes," AIAA Journal, Vol. 20, No. 11, November 1982, pp. 1572-1978.

10. Von Mises, R., "Mathematical Theory of Compressible Fluid Flow," Academic Press, Inc., New York, 1958, page 241.

11. South, J. and Jameson, A., unpublished NASA TP, 1979.

12. Ecer, A. and Akay, H. U., "Finite Element Analysis of Transonic Flows in Cascades - Importance of Computational Grids in Improving Accuracy and Convergence," NASA CR 3446, 1981.

13. Caspar, J. R., "Unconditionally Stable Calculation of Transonic Potential Flow Through Cascades Using an Adaptive Mesh for Shock Capture," ASME Paper No. 82-GT-238.

14. Sichel, M., "Structure of Weak Non-Hugoniot Shocks," The Physics of Fluids, Vol. 6, May 1963, pp. 653-663.

15. Chin, W., "Numerical Solution for Viscous Transonic Flow," AIAA Journal, Vol. 15, No. 9, pp. 1360-1362, September 1977.

16. Chin, W., "Algorithm for Inviscid Flow Using the Viscous Transonic Equation," AIAA Journal, Vol. 16, No. 8, pp. 848-849, August 1978.

17. Chin, W., "Type-Independent Solution for Mixed Compressible Flows," AIAA Journal, Vol. 16, No. 8, pp. 854-856, August 1978.

DIAGNOSTIC ALGORITHMS FOR CONTOUR DYNAMICS

Edward A. Overman II and Norman J. Zabusky
Institute for Computational Mathematics and Applications
Department of Mathematics and Statistics
University of Pittsburgh
Pittsburgh, PA 15261

ABSTRACT. The goal of large scale numerical simulations (numerical experiments) is to obtain a quantitative understanding of complicated nonlinear dynamical processes. A proper picture or graph can spark insights into new mathematical or physical processes and liberate us from the prejudices of our conservative intuitions. Diagnostic algorithms and their graphs are particularly useful in the contour dynamics model for studying two dimensional fluid dynamics. This is because the 2D densities are replaced by contours bounding piecewise-constant density regions (i.e., 1D curves). Thus our diagnostic parameters are functions of one variable, the arc length along each curve, and their graphs are 2D. We discuss and illustrate some time dependent properties of planar curves, including the spatial plot, low order moments, perimeter, curvature, and Fourier transforms. We will also apply these techniques to contours obtained from finite-difference representations of continuum systems.

I. INTRODUCTION. The method of contour dynamics is ideally suited to treating the dynamics of incompressible-inviscid fluids in two dimensions. For example, Longuet-Higgins and Cokelet [1,2] have studied incompressible shallow and deep water waves on boundaries between regions where the density is piecewise-constant. Zabusky, Hughes, Roberts, Deem, Overman, and Wu [3,4,5,6] have investigated the Euler equations with piecewise-constant finite-area-vortex-regions (FAVRs). Zabusky and Overman [7] have studied how to model surface tension and dissipation using contour dynamics. Finally, Overman, Zabusky, and Ossakow [8] have been studying the evolution of a piecewise-constant, weakly-ionized and strongly magnetized plasma in an electric field, a problem mathematically analogous to the Buckley-Leverett equations of flow in a porous media [9].

In this paper we discuss how diagnostics help to quantify and to gain insight into the time evolution of two very different fluid dynamical problems, namely the Euler equations and ionospheric plasma clouds. In Section II we discuss the two mathematical models and in Section III how they are discretized, including how we obtain the contour representations. In Section IV we will use various diagnostics and their graphs to study specific examples and in Section V we will show how these methods can be applied to continuum models.

II. MATHEMATICAL MODELS. The Euler equations can be written in vorticity-stream function form as

$$\omega_t + u\omega_x + v\omega_y = 0, \tag{1a}$$

where

$$\Delta\psi = -\omega, \tag{1b}$$

and

$$(u,v) \equiv (\psi_y, -\psi_x). \tag{1c}$$

If the vorticity is represented by a set of N_c piecewise-constant functions of strength ω_j in regions D_j with boundaries ∂D_j, we can express the stream function as

$$\psi(x,y) = \sum_{j=1}^{N_c} \omega_j \iint_{D_j} G(x-\xi, y-\eta)\,d\xi d\eta, \tag{2}$$

where G is the Green's function for the Laplacian in the unbounded domain

$$G(x-\xi, y-\eta) = -(2\pi)^{-1}\log[(x-\xi)^2 + (y-\eta)^2]^{1/2} = -(2\pi)^{-1} \log \ell. \tag{3}$$

If Green's theorem is applied to the result of substituting (2) into (1c) we obtain an expression for the velocity as a sum over the N_c contour integrals, namely

$$(u,v) \equiv (u(x,y), v(x,y)) = (2\pi)^{-1} \sum_{j=1}^{N_c} [\omega]_j \int_{\partial D_j} \log \ell(d\xi, d\eta), \tag{4}$$

where $[\omega]_j$ is the jump in vorticity (outside-inside) at ∂D_j and where the dependence on time has been suppressed. Alternately, if we integrate by parts we obtain

$$(u,v) = (2\pi)^{-1} \sum_{j=1}^{N_c} [\omega]_j \int_{\partial D_j} \ell^{-1}(x-\xi, y-\eta)\,d\ell. \tag{5}$$

The contours are advected by $(d/dt)(x,y) = (u,v)$.

The equations of motion of the ionospheric plasma cloud is [8]

$$\vec{\nabla} \cdot (N\vec{\nabla}\phi) = 0, \tag{6a}$$

$$\partial_t N + \vec{v}\cdot\vec{\nabla}N = \nu\vec{\nabla}^2 N, \tag{6b}$$

$$\vec{v} = (-\vec{e}_x\partial_y + \vec{e}_y\partial_x)\phi, \tag{6c}$$

where N is the density of ions, $\vec{E} = -\vec{\nabla}\phi$ is the electric field and $\vec{E} \rightarrow E_o\vec{e}_x$ as $|(x,y)| \rightarrow \infty$, and ν is the dissipation parameter arising from ion-neutral atom collisions. A two-dimensional problem arises because we assume that the ions move orthogonally to the magnetic field $B_o\vec{e}_z$ which is assumed to be large and unaffected by the cloud.

In the contour dynamical representation we assume that N is piecewise-constant and follow the evolution of smooth contours ∂D at which N is discontinuous. In the simplest case N takes on only two constant values, i.e.,

$$N(x,y,t) = \begin{cases} N_- & \text{for} \quad (x,y) \in D(t) \\ N_+ & \text{for} \quad (x,y) \notin D(t) \end{cases} \tag{7}$$

where $D(t)$ is a bounded, simply connected region in $\mathbb{R}^2$. Using the single- and double-layer Green's function of Laplace's equation, we find that [8]

$$\begin{aligned} \phi(p) &= c_1\, E_o x(p) \quad + c_2 \int_{\partial D} \phi(s)\partial_{n_s} \ln r(p,s)ds \\ \partial_n\phi(p) &= c_1\, E_o \partial_n x(p) + c_2 \int_{\partial D} \partial_n\phi(s)\partial_{n_p} \ln r(p,s)ds \end{aligned} \tag{8}$$

where

$$c_1 = -2N_+/(N_-+N_+), \qquad c_2 = -(1/\pi)(N_--N_+)/(N_-+N_+)$$

and where $s,p \in \partial D$, $r(p,s)$ is the straight line distance between p and s, and ∂_{n_s} and ∂_{n_p} are the derivatives in the direction of the outward normal to the boundary at s and p, respectively. Using Eq. 8 we obtain the electric field on ∂D. We then advect ∂D

by [8]

$$\frac{d}{dt}(x,y) = (-\partial_y\phi, \partial_x\phi) + \nu\,\partial_{ss}(x,y) \qquad (9)$$

where s is the arc length on ∂D.

III. DIAGNOSTICS. To discretize the contour, $\{(x,y) \in \partial D\}$, or contours we use a finite number of nodes, $\{(x_j,y_j) \mid 1 \le j \le N\}$, and follow the time evolution of these nodes. For the Euler equations we use the midpoint method to evaluate the integral in Eq. 5 and predictor-corrector to advect the nodes [6]. For the plasma cloud [8] we discretize Eqs. (8) by the trapezoidal method and use Gauss-Seidel to solve the resulting system of equations for the electric field at each node. The points are then advected, Eq. 9, using implicit predictor-corrector.

To obtain a continuum representation from the finite number of nodes we use cubic splines. That is [7], we calculate the straight line distance between adjacent nodes,

$$\Delta s_j = ((x_{j+1}-x_j)^2 + (y_{j+1}-y_j)^2)^{1/2},$$

and then calculate the cubic spline representations for $\{(s_j,x_j) \mid 1 \le j \le N\}$ and $\{(s_j,y_j) \mid 1 \le j \le N\}$. This continuum representation can then be used to [8] adjust points on the contour, calculate derivatives (e.g., to calculate the tangent angle and curvature), and calculate integrals (e.g., the area and center of mass).

IV. DIAGNOSTICS. In our study of the Euler equations one area of particular interest is steady-state solutions (V-states) and their stability. For example in [5b] we investigated the stability of a pair of rotating symmetric FAVRs (see Fig. 2). In Fig. 1 we show diagnostics for a V-state (calculated numerically) which was run for 125 units of time (slightly more than two revolutions). We show the area change, $\Delta A/A$, the perimeter change, $\Delta P/P$, the change in the distance between the center of rotation and the center of area, $\Delta\bar{x}/\bar{x}$, and the change in maximum curvature, $\Delta\kappa/\kappa$, for one of the contours. These diagnostics provide strong support that we have indeed found a steady-state solution. To investigate its stability we moved the contours away from the origin by 10%. In Fig. 2 we show the resulting spatial and curvature plots for nearly three revolutions. In Fig. 3 we show the corresponding diagnostics. Note that the perimeter, $\bar{x}$, and κ changes are a factor of 10 larger than in the steady-state case. Since the perimeter and $\bar{x}$ changes show no monotonic growth we conclude, numerically, that this configuration is stable. We also moved the contours in toward the origin by 2.5% from the steady-state solution and show the merger of the two FAVRs in Figs. 4 and 5. In the diagnostics we have replaced $\Delta\kappa/\kappa$ by $\ln|\Delta\bar{x}/\bar{x}|$ to show the rate of approach of the two contours more clearly. By use of these diagnostics we have numerically verified a conjecture by Saffman and Szeto [10] of the transition point between stability and instability for symmetric corotating V-states.

In our study of the plasma clouds there is an additional difficulty which must be considered. Namely, without dissipation (i.e., ν in Eq. 9) the contour dynamics model seems to be unstable or ill-posed [8]. (We believe the same is true of the continuum model, Eqs. 6.) Thus it is possible for round-off errors to cause large numerical errors in a very short time unless extreme care is taken. (This difficulty does not seem to occur in the Euler equations where it seems to be the low modes that grow in merger or fission [5] and there does not seem to be any high frequency instability.)

One of the basic questions we have addressed in plasma clouds is the effect of one cloud on another. We address this question in Fig. 6 where we show the time evolution of one cloud versus two clouds. Initially there is a 0.01% perturbation of the boundary at the 40th mode. As our diagnostics we use the curvature and the Fourier modes, i.e., we let the radius, r, be a function of the arc length, s, and decompose $r(s)$ into its Fourier modes. At time 0 you can see the perturbation at the 40th and also a small numerical error at the first mode due to the fact that we have distributed more points (there are 160 on each cloud) on the top of the contour, where it will go unstable, than on the bottom. From the comparison of the two cases we can see that the interaction of the two clouds causes a large change in the behavior of the low modes due to the clouds trying to move away from one another. However the high modes are hardly affected and we see the same number of striations ("fingers") forming on the top of the cloud. From numerical experiments such as these we have shown that the intermediate-time structure of a cloud is independent of the clouds surrounding it.

We have also used the curvature diagnostic to validate the accuracy of the numerical algorithm. We compared the evolution of a cloud with an initial perturbation of 1.0% at the 40th mode and two resolutions: (a) 400 nodes on the contour and a time step of 0.005; and (b) 560 nodes and a time step of 0.0025 (that is, 10 or 14 nodes per period, respectively). In Fig. 7 we show the curvatures at $t = 1.2$ where one can perceive only very slight differences. To see that the near-inflection points and new small-scale "wiggles" are not numerical artifacts we show the spatial evolution of the cloud in Fig. 8 and the curvature evolution in Fig. 9. These inflection points and "wiggles" show how the fingers are trying to align themselves vertically as they lengthen and thus they predict the spatial structure that will later appear.

V. CONTINUUM MODEL. It is possible to use these same kinds of diagnostics on two-dimensional continuum models by focusing on contour lines. We will restrict our attention to the plasma cloud example where all the runs were done using a finite-difference representation of Eqs. 6 developed by S. T. Zalesak at the Naval Research Laboratory [11]. We warn the reader that the length and time scales in the plots in this section are very different from those used previously and, in addition, the vertical scales in the diagnostics in this section vary from time to time.

In Fig. 10 at time 0 we show the initial density of the cloud as a contour plot. This initial condition was chosen to closely resemble the piecewise-constant clouds of contour dynamics and so the density is nearly constant inside and outside the contour lines and has a steep slope to approximate the jump in density in the contour dynamics clouds. There is an initial perturbation of 1.0% at the 12th mode. The contour line we will follow is the one at the average density of the cloud (that is, the average of the density at the center of the cloud and at infinity at time 0). The diagnostics we will use are the curvature, the tangent angle, and the density gradient perpendicular to the contour (i.e., the maximum density gradient). The diagnostics are all smoothed by least squares interpolation to remove high frequency noise since, for example, the small-scale fluctuations in the curvature can be up to 3 times as large as the actual curvature due to the calculation of the contour points from the 2D mesh. At time 0 the curvature and tangent angle are very close to their analytical values but the gradient, which should be constant, varies by 20%. This is due to the fact that a variable mesh was used and the mesh size was very large at the bottom of the cloud, which is stable. (This effect can also be seen in the curvature at the bottom of the cloud at later times.) The gradient diagnostic is particularly valuable in plasma clouds because it is the large gradient which causes the fingers to form and, at later times, secondary fingers to split off.

ACKNOWLEDGMENTS. This work has been supported by the Army Research Office, the Office of Naval Research, and the Defense Nuclear Agency.

REFERENCES.

1. M. S. Longuet-Higgins and E. D. Cokelet, Proc. Roy. Soc. A 350, 1 (1976). M. S. Longuet-Higgins and E. D. Cokelet, Proc. Roy. Soc. A 364, 1 (1978).

2. M. S. Longuet-Higgins, J. Fluid Mech. 107, 1 (1981).

3. N. J. Zabusky. M. H. Hughes, and K. V. Roberts, *J. Comp. Phys. 30*, 96 (1979).

4. G. S. Deem and N. J. Zabusky, *Phys. Rev. Lett. 40*, 859 (1978).

5. E. A. Overman, II and N. J. Zabusky, *J. Fluid Mech. 125*, 187 (1982).
E. A. Overman, II and N. J. Zabusky, *Phys. Fluids 25*, 1297 (1982).

6. H. M. Wu, E. A. Overman, II and N. J. Zabusky, to appear in J. Comput. Phys.

7. N. J. Zabusky and E. A. Overman, II, to appear in J. Comput. Phys.

8. E. A. Overman, II, N. J. Zabusky, and S. L. Ossakow, *Phys. Fluids 26*, 1139 (1983).

9. D. W. Peaceman, "Fundamentals of Numerical Reservoir Simulation", pp. 19-22, Elsevier-Amsterdam (1977).

10. P. G. Saffman and R. Szeto, *Phys. Fluids 23*, 2339 (1980).

11. S. T. Zalesak, NRL Memorandum Report 5119 (1983).

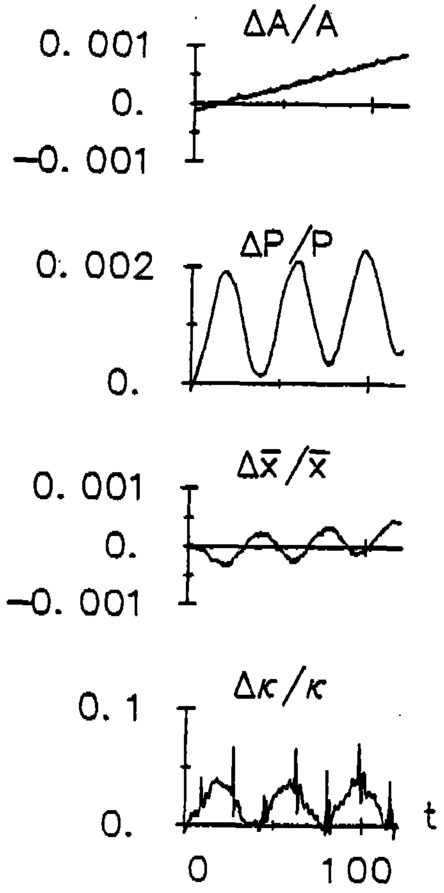


Figure 1

Diagnostics for the unperturbed V-state: change in area, $\Delta A/A \equiv (A(t)-A(0))/A(0)$; change in perimeter, $\Delta P/P \equiv (P(t)-P(0))/P(0)$; change in center of mass, $\Delta\bar{x}/\bar{x} \equiv (\bar{x}(t)-\bar{x}(0))/\bar{x}(0)$; and, change in maximum curvature, $\Delta\kappa/\kappa \equiv (\max|\kappa(t)|-\max|\kappa(0)|)/\max|\kappa(0)|$. (Fig. 2 of Ref. 5b.)

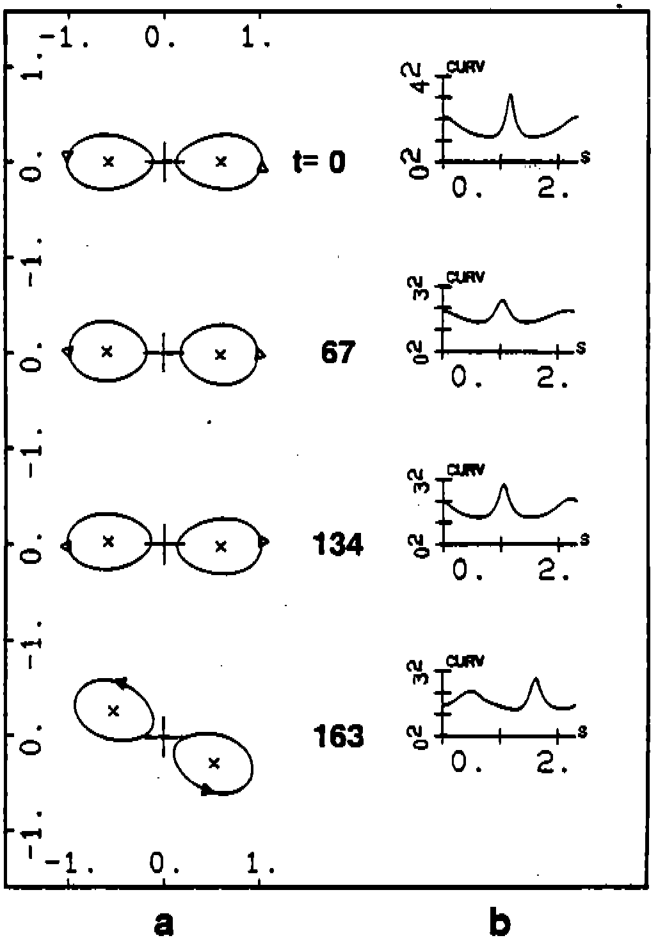


Figure 2

Outwardly perturbed V-state: (a) physical space; (b) curvature vs. arc length. (Fig. 3 of Ref. 5b.)

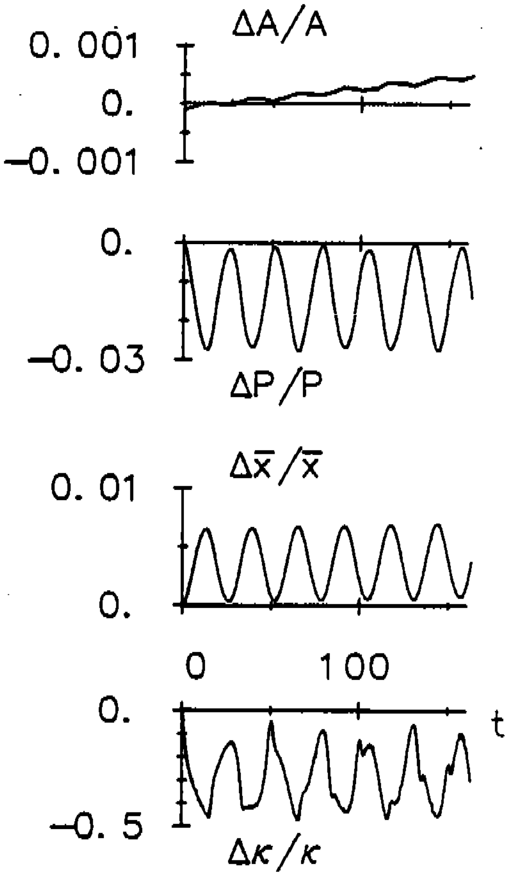


Figure 3

Outwardly perturbed V-state. The plots are the same as in Fig. 1. (Fig. 4 of Ref. 5b.)

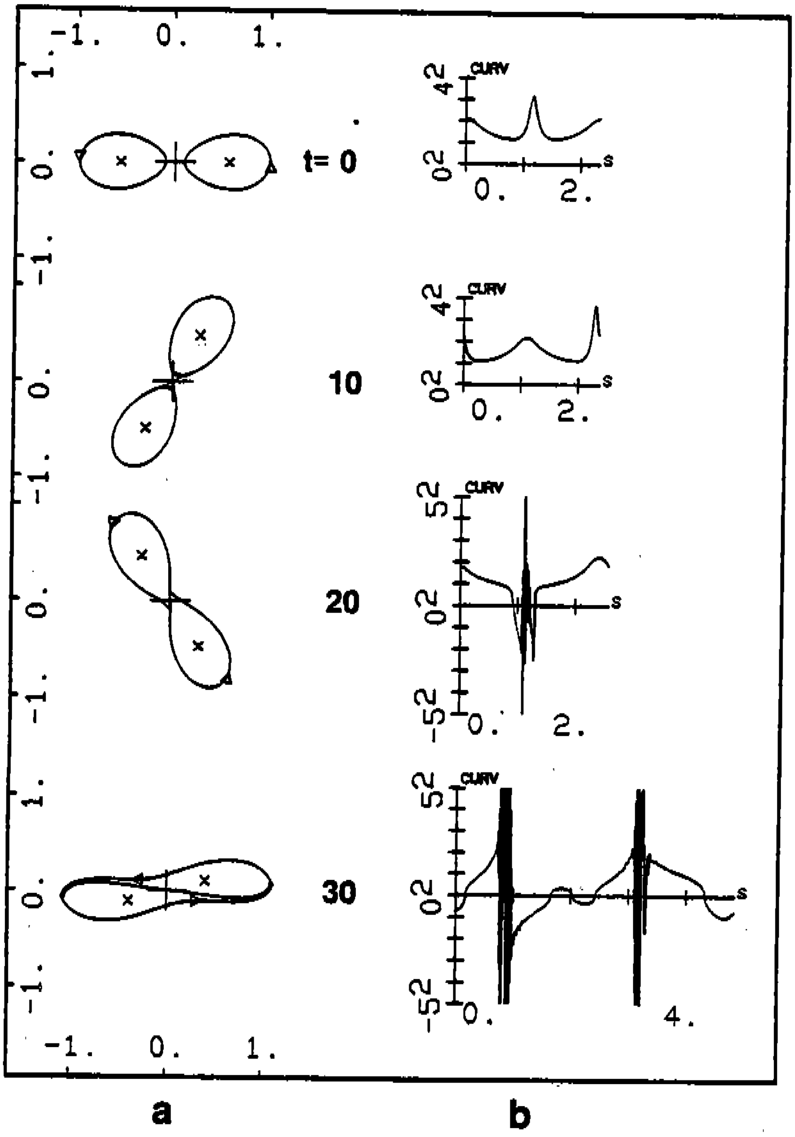


Figure 4

Inwardly perturbed V-state. The plots are the same as Fig. 2. (Fig. 5 of Ref. 5b.)

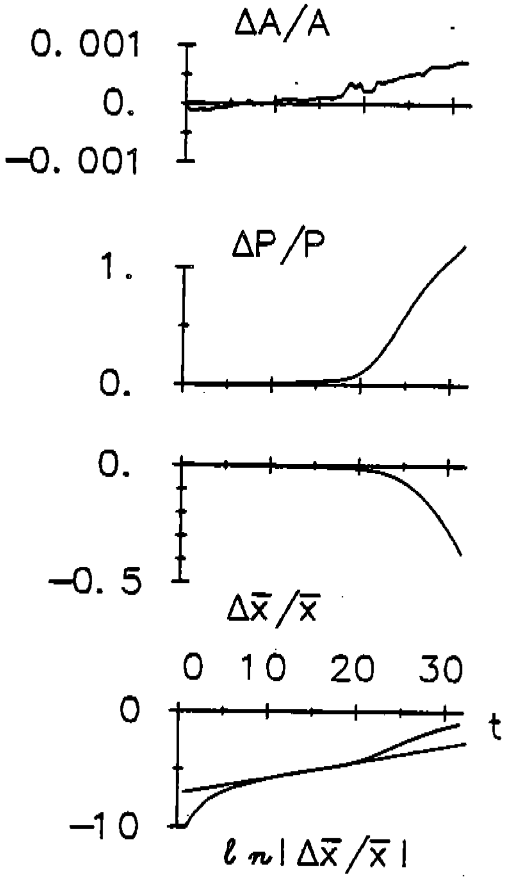


Figure 5

Inwardly perturbed V-state. The plots are the same as in Fig. 1 except that the change in curvature has been replaced by the logarithm of the change in center of mass. (Fig. 6 of Ref. 5b.)

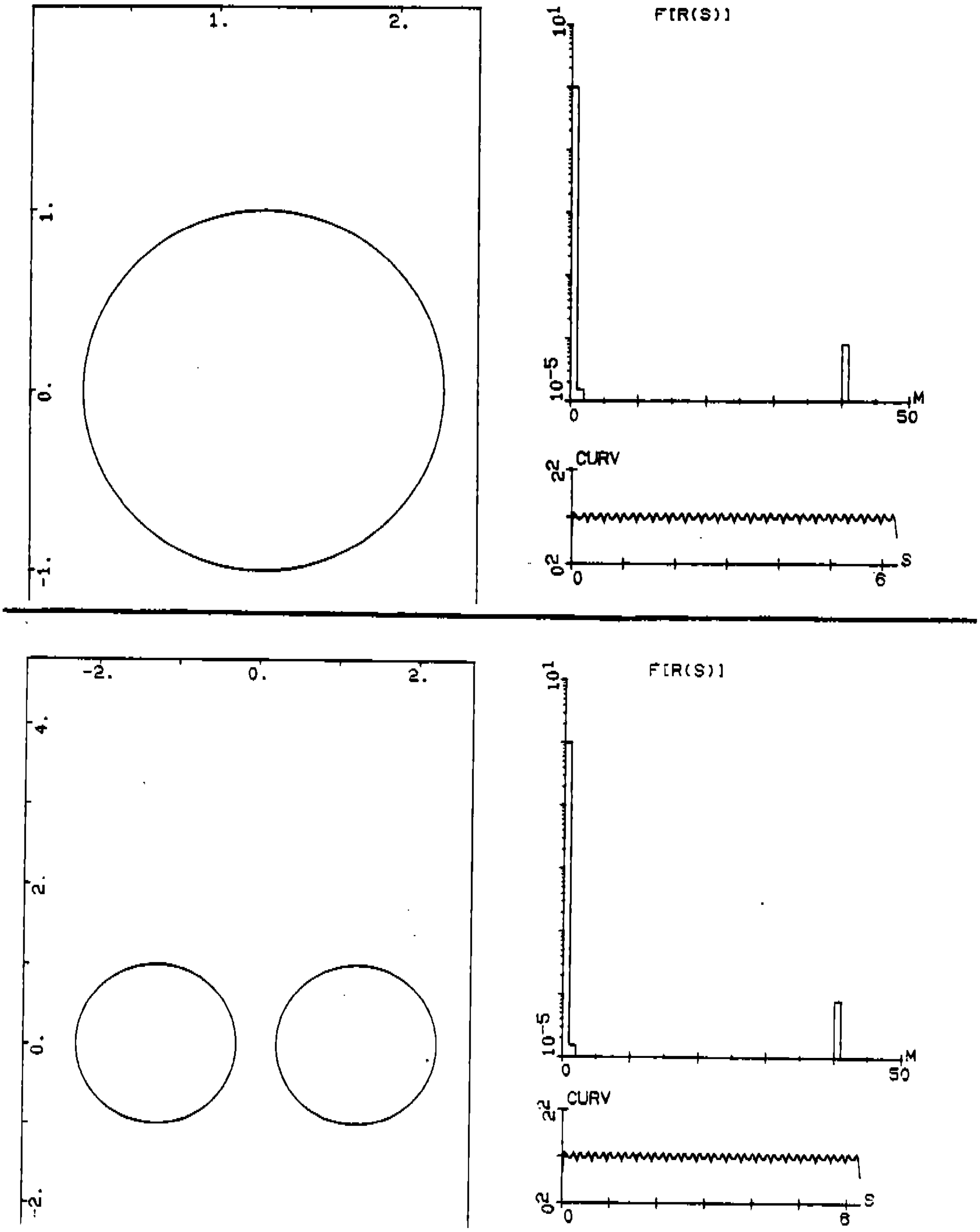


Figure 6

Time = 0. A comparison of the time evolution of one and two clouds with a 0.01% perturbation at the 40th mode. Shown are: physical space, the curvature vs. arc length, s (s = 0 is the bottom of the cloud), and, the amplitude of the Fourier modes, M, of r(s) where r is the radius from the center of mass.

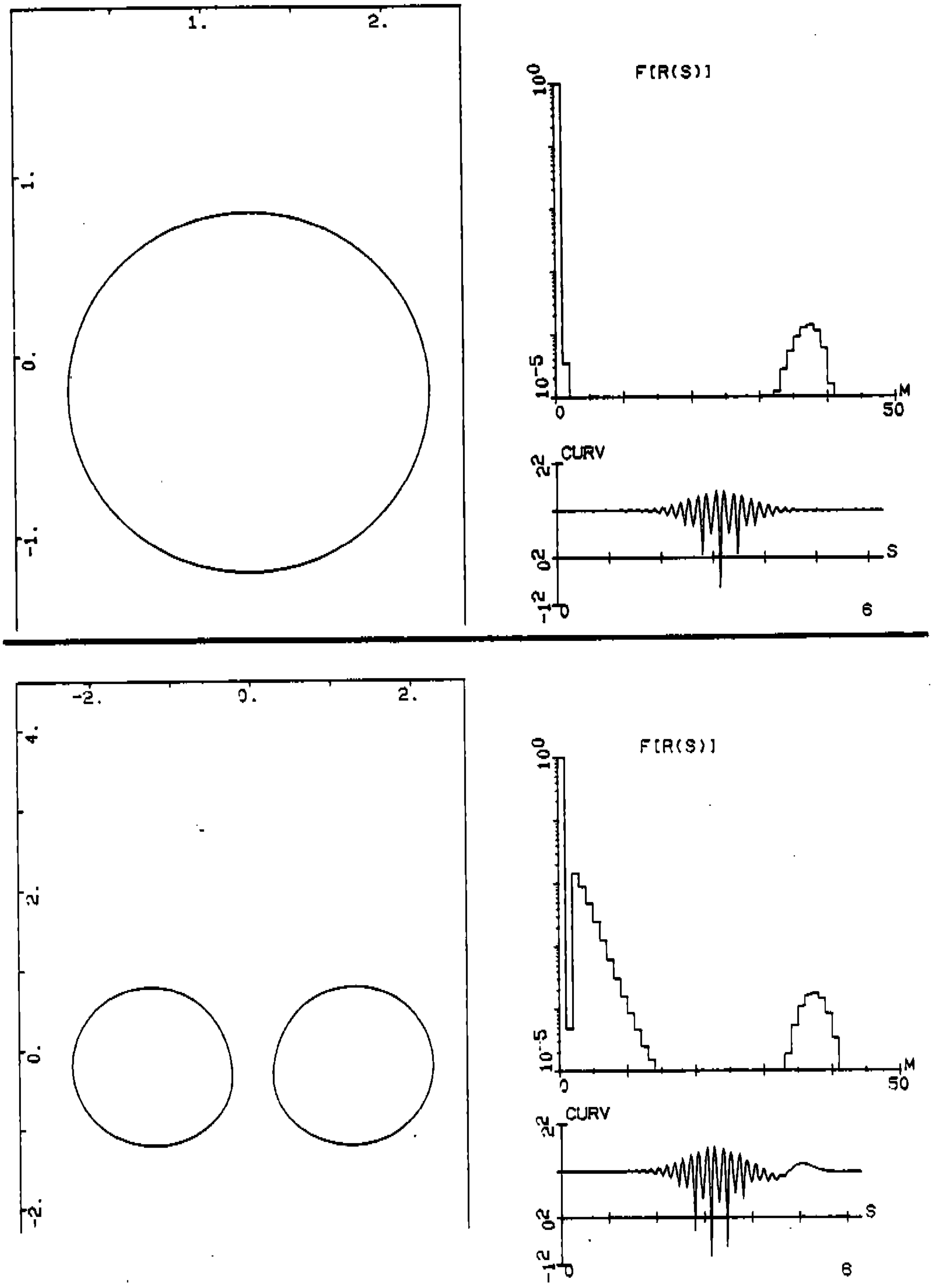


Figure 6 (continued). Time = 0.40.

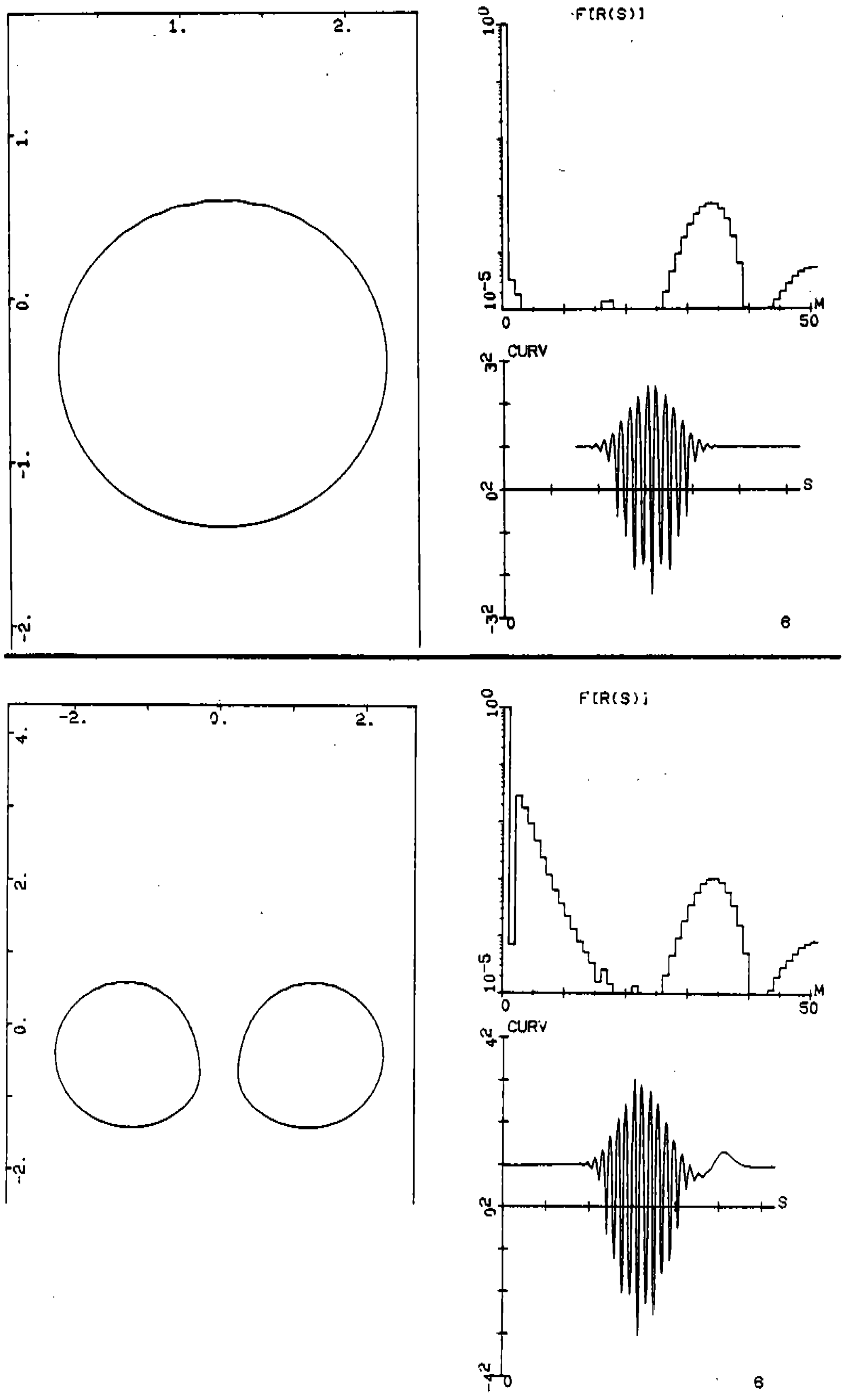


Figure 6 (continued). Time = 0.80.

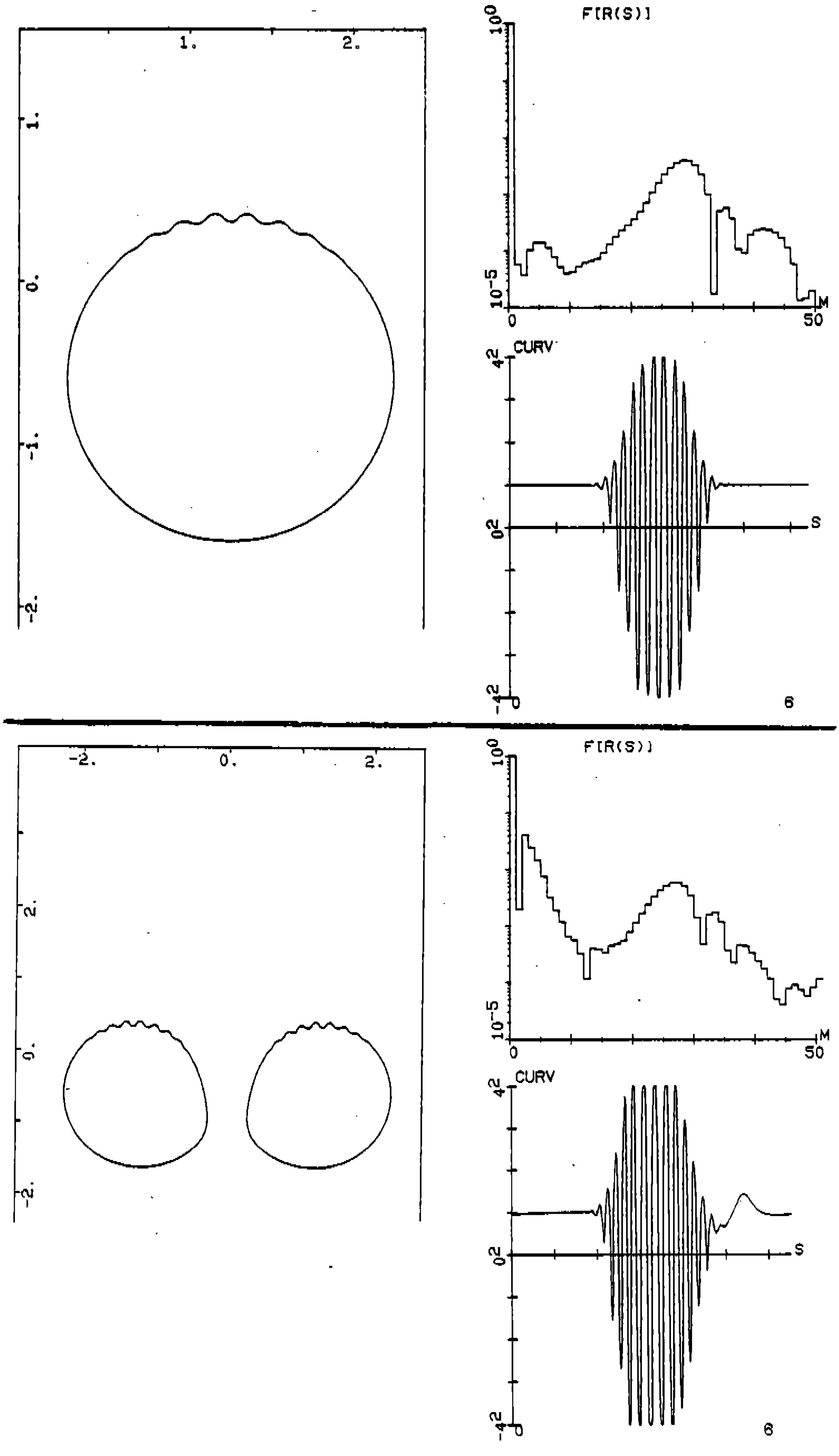


Figure 6 (continued). Time = 1.20.

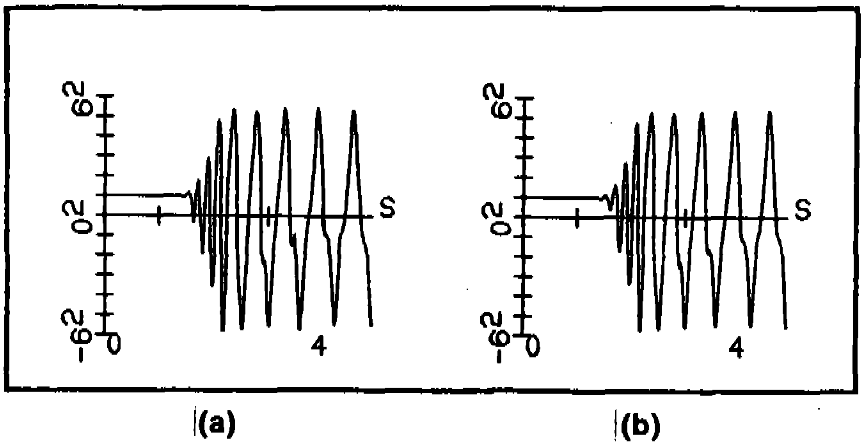


Figure 7

Validation of the accuracy of the contour dynamics code. The evolution of a cloud with an initial perturbation of 1% at the 40th mode and two resolutions: (a) 400 nodes and Δt = 0.005; and (b) 560 nodes and Δt = 0.0025. The curvatures vs. arc length are shown at time = 1.20. (s = 0 is the bottom of the cloud and s = maximum is the top.) (Fig. 10 of Ref. 8.)

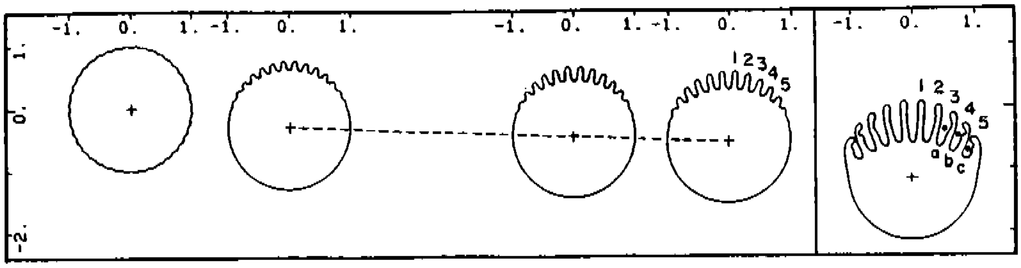


Figure 8

The spatial evolution for the same run as in Fig. 7. The times, from left to right, are 0.00, 0.60, 0.90, 1.10 and 2.30. (Fig. 11 of Ref. 8.)

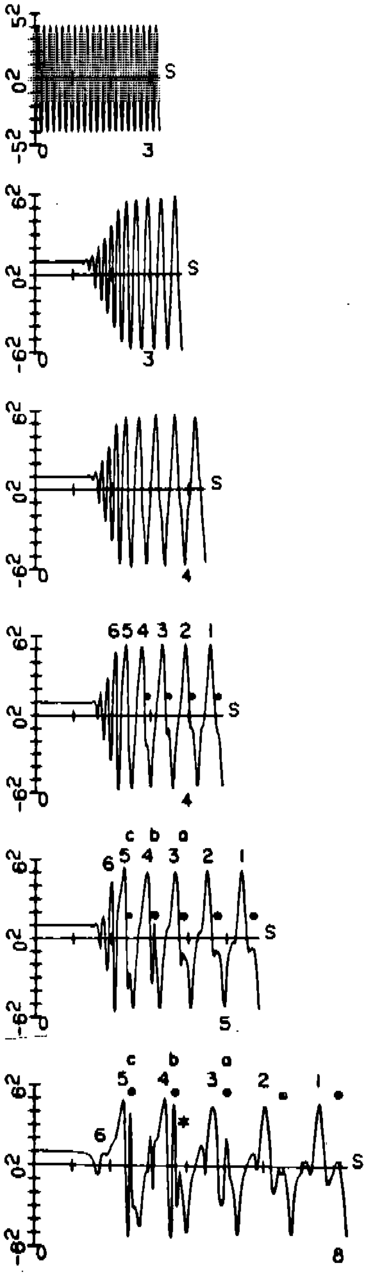


Figure 9

The curvature vs. arc length corresponding to Fig. 8. (Fig. 12 of Ref. 8.)

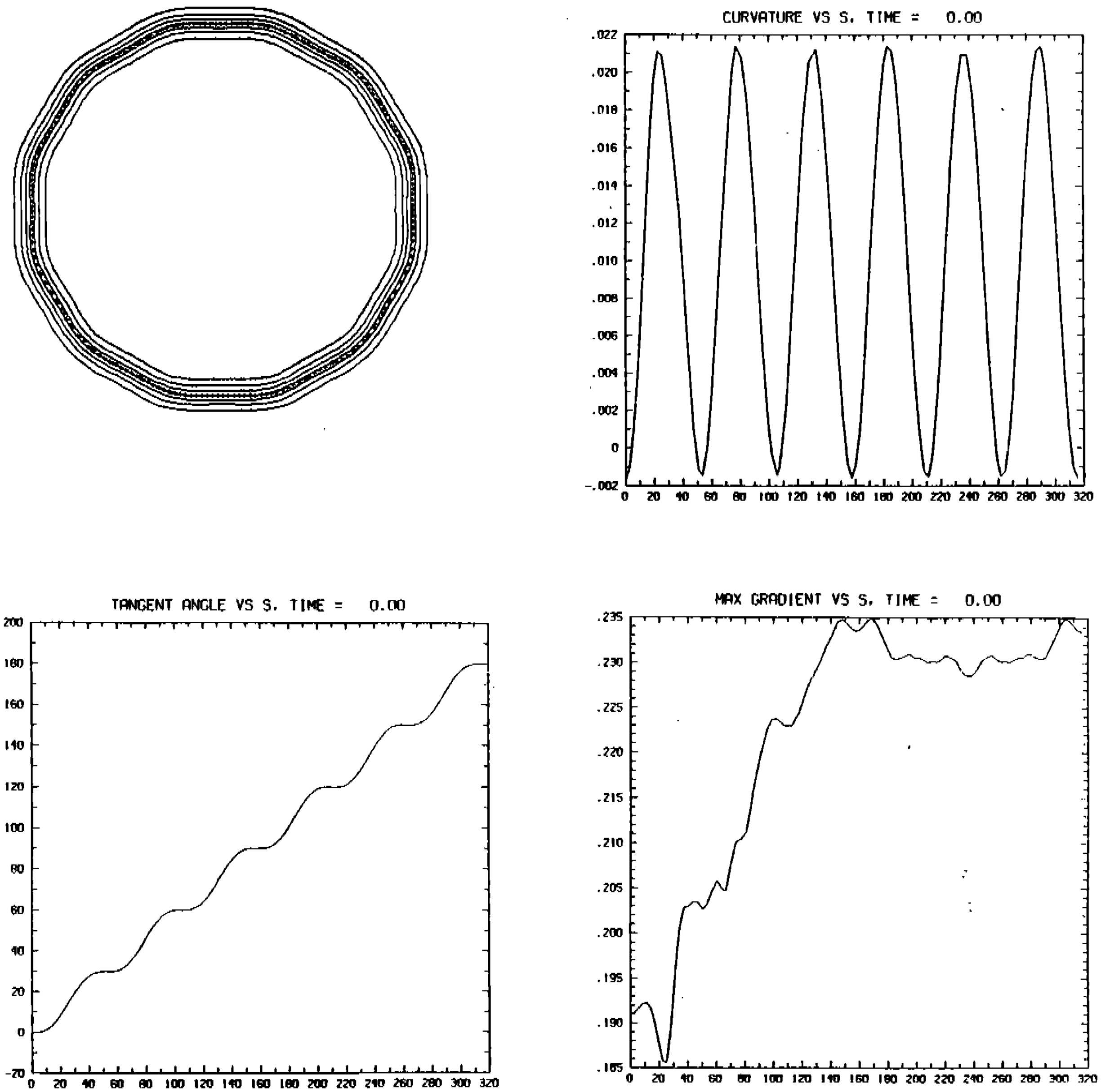


Figure 10

Time = 0. The time evolution of a run using the continuum equations, Eqs. 6. Shown are: a contour plot of the physical space cloud with the middle contour used in the diagnostic plots; the curvature vs. arc length (s = 0 at the bottom and s = maximum at the top); the tangent angle vs. arc length; and, the maximum gradient (i.e., the density gradient perpendicular to the contour) vs. arc length.

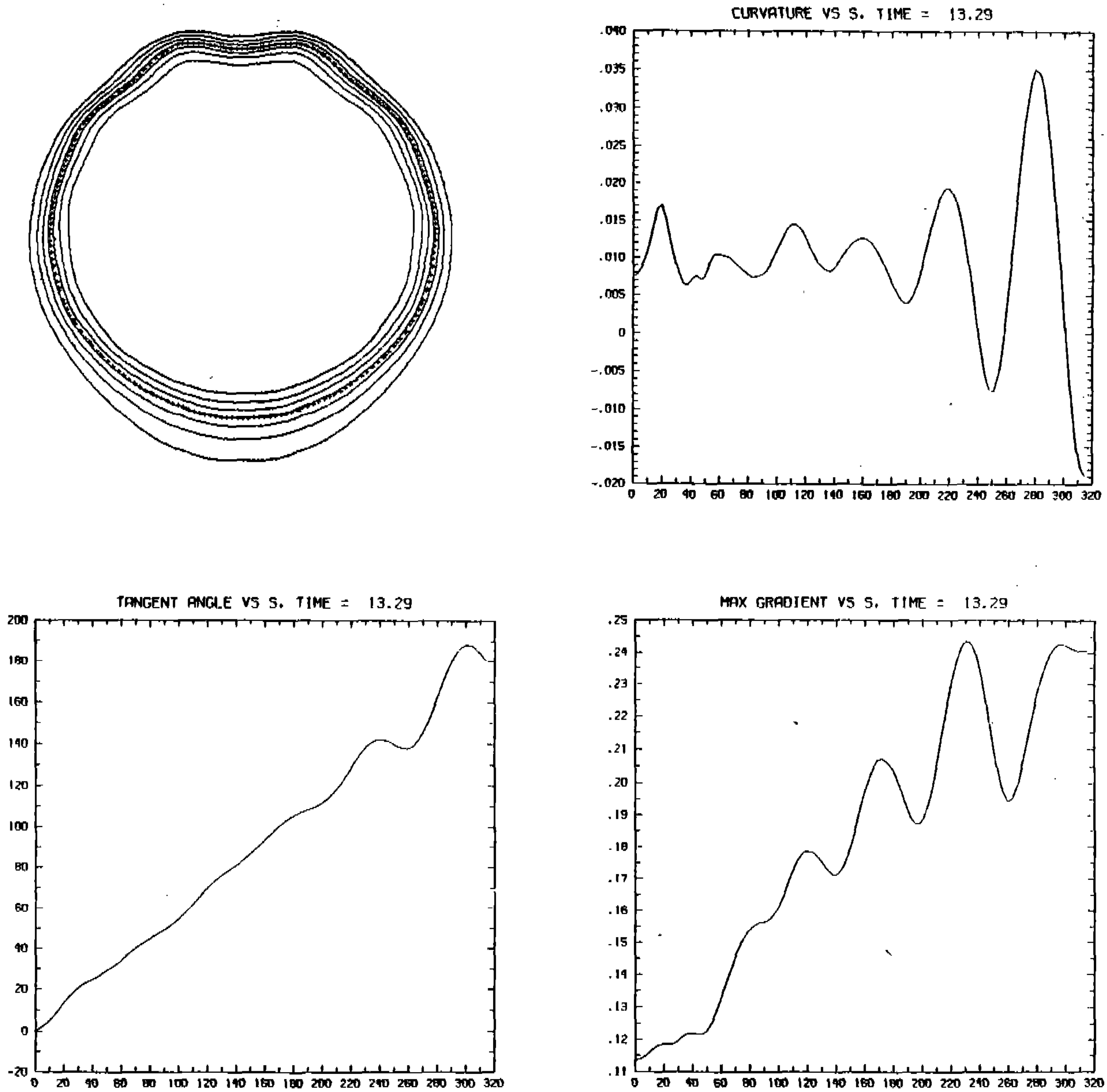


Figure 10 (continued). Time = 13.29.

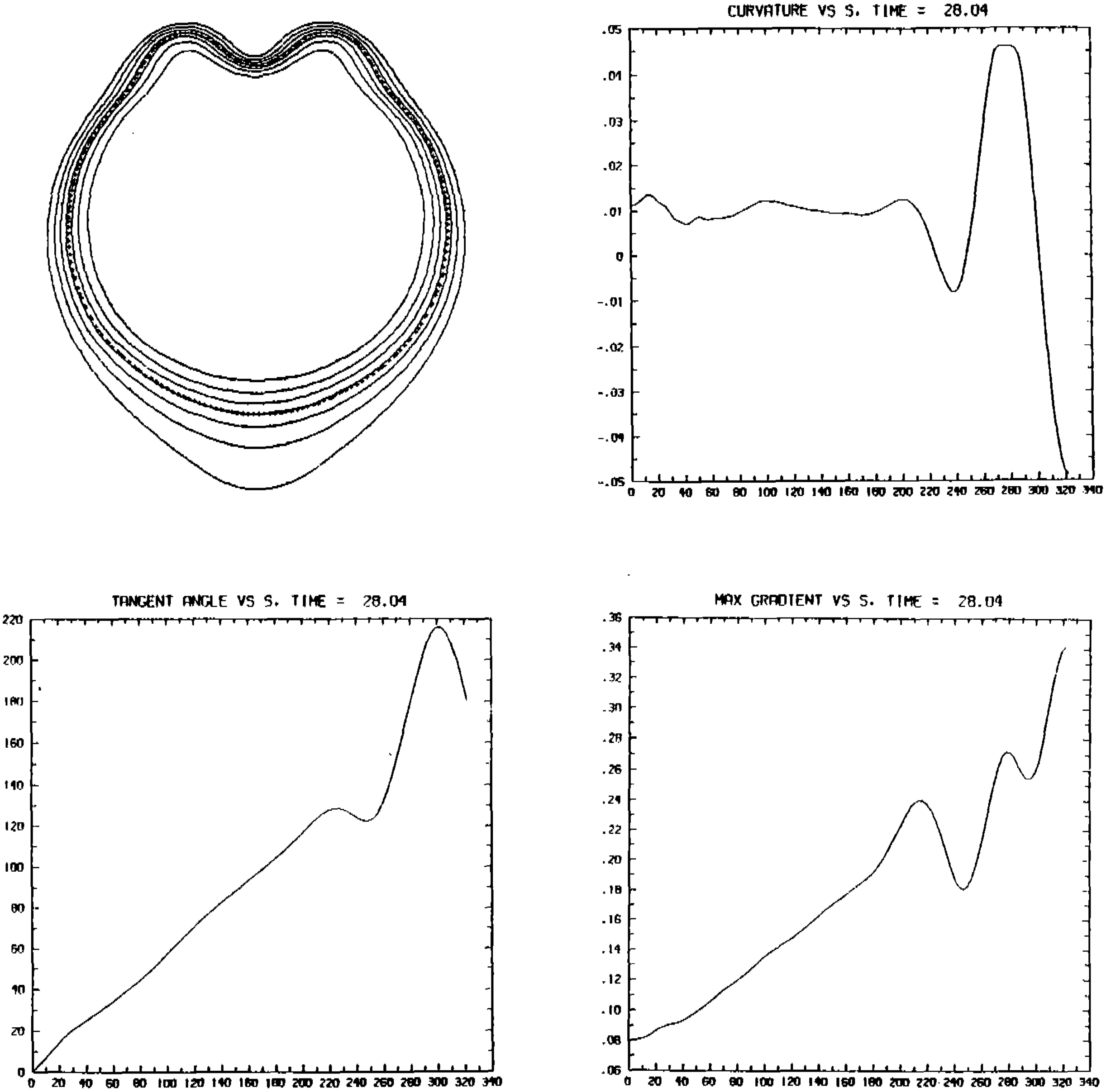


Figure 10 (continued). Time = 28.04.

ANALYSIS OF THE VON KARMAN EQUATIONS BY GROUP METHODS

K. A. Ames
Department of Mathematics
Iowa State University
Ames, Iowa 50011

W. F. Ames*
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT. One of the system of equations approximating the large deflection of plates consists of two coupled nonlinear fourth order partial differential equations, known as the von Karman equations. The full symmetry group for the steady equations is a finitely generated Lie group with ten parameters. For the time dependent system the full symmetry group is an infinite parameter Lie group. Several subgroups of the full group are used to generate exact solutions of the time-independent and the time-dependent system. These include the dilatation group (similar solutions), rotation group, screw group and others. Physical implications and applications are discussed.

0. INTRODUCTION. Perhaps the most widely applicable method for determining analytic solutions of partial differential equations utilize the underlying (Lie) group structure. The mathematical foundations for the determination of the full group for a system of differential equations can be found in Ames [1], Bluman and Cole [2], and the general theory is found in Ovsiannikov [3]. The determination of the full group requires extremely lengthy calculations. Detailed calculations can be found in Ames [1], Ovsiannikov [3] and for the Navier Stokes equations in Boisvert [4] (see also Boisvert, et al. [5]). Here we give the results of those calculations for the von Karman equations (see (0)) of nonlinear elasticity in the form of the infinitesimal generators of the full group. These have also been obtained in an independent study by Schwarz [6] using an algebraic program package which uses REDUCE [7]. In Russia such an algebraic programming system, for this purpose, is available under the name CINO (see Ovsiannikov [3], p. 57). MACZYMA (Roseneau and Schwarzmeier [8]) has also been used for this purpose on other problems.

Our goal is to obtain explicit invariant solutions to the system of partial differential equations, due to von Karman,

$$\begin{aligned} \Delta^2 F &= E[w_{xy}^2 - w_{xx}w_{yy}] \\ \Delta^2 w &= \frac{q}{D} + \frac{h^*}{D}[F_{yy}w_{xx} + F_{xx}w_{yy} - 2F_{xy}w_{xy}] \end{aligned} \tag{0}$$

*Research supported by U.S. Army Grant DAAG-29-81-K-0042.

by employing various subgroups of the full group admitted by these equations. To apply this procedure we choose a one (or more) parameter subgroup and calculate the general form of the subgroup invariants. We then require the equations to be invariant under this group. As a result a set of simultaneous algebraic equations arise and their solution, possibly involving arbitrary parameters, leads to a more specific form of the invariants. Substitution of these invariants into (0) results in a representation of the system from which solutions to the original system can be constructed.

A preliminary version of this paper appeared in [9].

1. THE EQUATIONS. The investigation of large deflections of plates rests on the solution of two coupled nonlinear partial differential equations known as the von Karman equations [10]. Let us consider a rectangular elastic plate under the combined action of a uniform lateral load and a tensile force in the middle plane of the plate. Denote by w the deflection of the plate away from its equilibrium position in a region D of the xy plane and by F the Airy stress function. Assuming that there are no body forces in the plane of the plate and that the lateral load is perpendicular to the plate, then w and F satisfy the equations

$$\Delta^2 F = E[(\frac{\partial^2 w}{\partial x \partial y})^2 - \frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}]; \tag{1.1}$$

$$\Delta^2 w = \frac{q}{D} + \frac{h^*}{D}[\frac{\partial^2 F}{\partial y^2}\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 F}{\partial x^2}\frac{\partial^2 w}{\partial y^2} - 2\frac{\partial^2 F}{\partial x \partial y}\frac{\partial^2 w}{\partial x \partial y}]. \tag{1.2}$$

Here E is the modulus of elasticity, D is the flexural rigidity, h* is the thickness of the plate, q is the lateral load intensity and Δ^2 is the biharmonic operator, i.e.,

$$\Delta^2 \equiv \frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}$$

Determination of the stress function F allows us to calculate the stresses in the middle surface of the plate by means of the relations

$$\sigma_x = \frac{\partial^2 F}{\partial y^2}; \quad \sigma_y = \frac{\partial^2 F}{\partial x^2}; \quad \tau_{xy} = -\frac{\partial^2 F}{\partial x \partial y} \tag{1.3}$$

From the function w, which defines the deflection surface of the plate, we can obtain the bending and shearing stresses.

In our group analysis of equations (1.1) and (1.2), we shall be interested in the two cases $q = 0$ and $q = -\rho\partial^2 w/\partial t^2$. The first of these cases represents the situation in which there is no lateral loading while the second describes the vibration of a plate whose deflections are large in comparison with its thickness. Rather than treat (1.1) and (1.2) directly, we introduce the new variables, $\bar{x}$, $\bar{y}$, $\bar{t}$, $\bar{F}$ and $\bar{w}$ which are defined by

$$x = m\bar{x}, \quad y = m\bar{y}, \quad t = p\bar{t}, \quad F = b\bar{F}, \quad w = c\bar{w}.$$

The choices

$$b = \frac{D}{h^*} \text{ and } c^2 = \frac{D}{(h^*E)}$$

as well as

$$m = D^{1/2} \text{ and } p^2 = \rho(\frac{h^*D}{E})^{1/2} ,$$

in the case $q = -\rho\partial^2 w/\partial t^2$, lead to the dimensionless sytem

$$\Delta^2 F = (\frac{\partial^2 w}{\partial x \partial y})^2 - \frac{\partial^2 w}{\partial x^2} \frac{\partial^2 w}{\partial y^2} \tag{1.4}$$

$$\Delta^2 w = \frac{\partial^2 F}{\partial y^2} \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 F}{\partial x^2} \frac{\partial^2 w}{\partial y^2} - 2 \frac{\partial^2 F}{\partial x \partial y} \frac{\partial^2 w}{\partial x \partial y} \tag{1.5a}$$

or

$$\Delta^2 w = - \frac{\partial^2 w}{\partial t^2} + \frac{\partial^2 F}{\partial y^2} \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 F}{\partial x^2} \frac{\partial^2 w}{\partial y^2} - 2 \frac{\partial^2 F}{\partial x \partial y} \frac{\partial^2 w}{\partial x \partial y} \tag{1.5b}$$

where the bars have been dropped. We shall investigate (1.4)-(1.5a) or (1.4)-(1.5b) using group analytic techniques with the aim of obtaining exact solutions of these systems.

2. FULL GROUP. The full (Lie) group of the time-dependent von Karman equations (1.4, 1.5b) is given by 14 infinitesimal operators

$$\begin{gathered}
X_1 = \frac{\partial}{\partial t} ; \quad X_2 = \frac{\partial}{\partial x} ; \quad X_3 = \frac{\partial}{\partial y} ; \quad X_4 = y \frac{\partial}{\partial x} - x \frac{\partial}{\partial y} ; \\
X_5 = 2t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} ; \quad X_6 = \frac{\partial}{\partial w} ; \quad X_7 = t \frac{\partial}{\partial w} ; \\
X_8 = x \frac{\partial}{\partial w} ; \quad X_9 = y \frac{\partial}{\partial w} ; \quad X_{10} = tx \frac{\partial}{\partial w} ; \quad X_{11} = ty \frac{\partial}{\partial w} ; \\
X_{12} = f_1(t) x \frac{\partial}{\partial F} ; \quad X_{13} = f_2(t) y \frac{\partial}{\partial F} ; \quad X_{14} = f_3(t) \frac{\partial}{\partial F} .
\end{gathered} \tag{2.1}$$

Each of the first 11 generators is associated with a parameter independent of all the others. These generate a finite dimensional Lie algebra L_{11}. The last three operators contain arbitrary functions of time, $f_i(t)$, $i = 1,2,3$.

3. DILATATION GROUP. In this section we investigate the action of the dilatation group (from X_5)

$$\bar{x} = a^{\alpha}x, \quad \bar{y} = a^{\alpha}y, \qquad (a > 0) \tag{3.1}$$

on the dimensionless time independent von Karman equations (1.4)-(1.5a). The invariants of this transformation group are ([1])

$$\eta = x/y. \tag{3.2}$$

The von Karman equations are invariant under the dilatation group and

$$F = f(\eta); \quad w = h(\eta). \tag{3.3}$$

Substitution of these relations into equations (1.4)-(1.5a) leads to the following two coupled ordinary differential equations for f and h:

$$\frac{d^3}{d\eta^3}\left[(1 + \eta^2 + \eta^4)\,\frac{df}{d\eta}\right] = \left(\frac{dh}{d\eta}\right)^2 ; \tag{3.4}$$

$$\frac{d^3}{d\eta^3}\left[(1 + \eta^2 + \eta^4)\,\frac{dh}{d\eta}\right] = -\left[\frac{d}{d\eta}\left(\eta\,\frac{df}{d\eta}\,\frac{dh}{d\eta}\right) + \frac{df}{d\eta}\,\frac{dh}{d\eta}\right]. \tag{3.5}$$

Several exact solutions of the system (3.4)-(3.5) can be obtained. If we choose the deflection to be constant (i.e., $h(\eta)$ = constant), then equation (3.5) is satisfied identically while (3.4) becomes

$$\frac{d^3}{d\eta^3}\left[(1 + \eta^2 + \eta^4)\,\frac{df}{d\eta}\right] = 0. \tag{3.6}$$

Integration of this equation gives the invariant solution

$$\begin{aligned} f(\eta) = \frac{1}{\sqrt{3}} \sum_{\nu=0}^{1} \big[&(c_1 \sin\phi_\nu)\,\ell n(z^2 - 2z\cos\phi_\nu + 1) \\ &+ (c_2 \cos\phi_\nu)\arctan[(z - \cos\phi_\nu)/\sin\phi_\nu] \\ &+ \frac{2c_3}{3}\arctan[(4\eta + 2)/3] + c_4, \end{aligned} \tag{3.7}$$

where $z = \sqrt{\eta}$, $\phi_\nu = -2\pi/3 + \nu\pi$ and the c_i $(i = 1,\ldots4)$ are arbitrary constants.

Another exact solution can be determined by letting $h(\eta) = cf(\eta)$ for a constant $c \neq 0$. Combining equations (3.4) and (3.5) we find that $f(\eta)$ must satisfy

$$\frac{d}{d\eta}\left[\eta\left(\frac{df}{d\eta}\right)^2\right] = -(1 + c^2)\left(\frac{df}{d\eta}\right)^2.$$

Thus

$$f(\eta) = \pm\,\frac{2B}{c^2}\,\eta^{-(c^2/2)} + D$$

$$h(\eta) = \pm \frac{2B}{c} \eta^{-(c^2/2)} + cD$$

where B and D are constants.

We note here that under the dilatation group the stresses acting in the middle plane of the plate are related by the equation $y^2\sigma_y + x^2\sigma_x = 2xy\tau_{xy}$. In practice, one is usually more interested in these quantities than the deflection surface.

Remark 2.1. Since the equations for the Airy stress function F and the deflection w are invariant under coordinate translations $\bar{x} = x + b$ and $\bar{y} = y + d$, it follows that the von Karman equations are also invariant under the group given by $\bar{\eta} = \bar{x}/\bar{y}$, $f(\bar{\eta})$ and $h(\bar{\eta})$.

4. ROTATION GROUP. We now examine the properties of the von Karman equations under the rotation group which, for any real a > 0, is defined by the transformation (from X_4)

$$\bar{x} = x \cos a - y \sin a; \quad \bar{y} = x \sin a + y \cos a; \quad \bar{w} = w; \quad \bar{F} = F.$$

It is known [1] that the invariants of this group are

$$\eta = x^2 + y^2; \quad F = f(\eta); \quad w = h(\eta). \tag{4.1}$$

Upon substituting (4.1) into equations (1.4)-(1.5a), we find that f and h must satisfy the two differential equations

$$4 \frac{d^2}{d\eta^2} [\eta^2 \frac{d^2f}{d\eta^2}] = - \frac{d}{d\eta}[\eta(\frac{dh}{d\eta})^2]; \tag{4.2}$$

$$2 \frac{d^2}{d\eta^2} [\eta^2 \frac{d^2h}{d\eta^2}] = \frac{d}{d\eta}[\eta \frac{dh}{d\eta} \frac{df}{d\eta}]. \tag{4.3}$$

Both of these equations can be integrated once to obtain

$$4 \frac{d}{d\eta} [\eta^2 \frac{d^2f}{d\eta^2}] = -\eta(\frac{dh}{d\eta})^2 + A; \tag{4.4}$$

$$2 \frac{d}{d\eta} [\eta^2 \frac{d^2h}{d\eta^2}] = \eta \frac{dh}{d\eta} \frac{df}{d\eta} + B, \tag{4.5}$$

for constants A and B. Two particular choices of h lead to exact solutions of this system. With h = constant, equation (4.5) is satisfied identically if we take B = 0 while (4.4) reduces to a linear equation which has solutions of the form (singular at the origin),

$$f(\eta) = c_1 + c_2 \ln \eta + c_3\eta + c_4\eta \ln \eta, \qquad (\eta \neq 0)$$

where the c_i are constants. A second approach is to set $h(\eta) = cf(\eta)$ for a constant c. Equations (4.4)-(4.5) then lead to

$$\eta(\frac{df}{d\eta})^2 = D^2 , \tag{4.6}$$

where the constant D depends upon A, B, and c. For $\eta \neq 0$, (4.6) integrates to give

$$f = \pm 2D\eta^{1/2} + \kappa,$$

with κ a constant of integration.

It is interesting to observe that if we set $dh/d\eta = df/d\eta$ in (4.4) or $df/d\eta = -dh/d\eta$ in (4.5), we obtain in both cases an equation of the form

$$\frac{d}{d\eta} [\eta^2 \frac{dg}{d\eta}] + \lambda\eta g^2 = c, \tag{4.7}$$

where $g = df/d\eta$ and $\lambda = 1/4$ or $g = dh/d\eta$ and $\lambda = 1/2$. For $c = 0$, (4.7) is an equation of the Emden-Fowler type about which there is considerable literature (e.g., see [11]). Because analytic solutions of this class of equations are known only in a few cases, the appearance of such an equation in the analysis is indicative of the difficulties one finds in attempting to find exact solutions of the von Karman equations. We note here that for $c = 0$, equation (4.7) can be transformed into

$$\frac{d^2y}{d\eta^2} + \eta^{-2}y^2 = 0 \tag{4.8}$$

by setting $y(\eta) = \eta\lambda g(\eta)$. One established fact about equation (4.8) is that all of its solutions are oscillatory on $(0,\infty)$, by which we mean that for any $\eta_1 > 0$ there exists $\eta_2 > \eta_1$ such that $y(\eta_2) = 0$. This type of result gives us information about a subset of those solutions of the von Karman equations which are invariant under the rotation transformation group.

5. SPIRAL GROUP. As we shall see, the von Karman equations also remain invariant under a third group of transformations, the spiral group, which has the form (a linear combination of X_2, X_3, X_6 and X_{14} (with $f_3(t)$ constant))

$$\bar{x} = x + \gamma_1 a, \quad \bar{y} = y + \gamma_2 a, \quad \bar{F} = e^{\alpha a}F, \quad \bar{w} = e^{\beta a}w, \tag{5.1}$$

for real $a > 0$ and γ_1, γ_2, α, β constants. The invariants of this group are

$$\eta = \gamma_2 x - \gamma_1 y, \quad f(\eta) = \gamma_2 \ln F - \alpha y, \quad h(\eta) = \gamma_2 \ln w - \beta y. \tag{5.2}$$

Substitution of (5.1) into equations (1.4)-(1.5a) leads to the two invariance conditions $\alpha = 2\beta$ and $\beta = \alpha+\beta$ which together imply that

$\alpha = \beta = 0$. Thus, (5.2) becomes

$$\eta = \gamma_2 x - \gamma_1 y, \quad f(\eta) = \gamma_2 \ln F, \quad h(\eta) = \gamma_2 \ln w \tag{5.3}$$

and the von Karman equations are reduced to a fourth order orindary differential equation which both f and h must satisfy, namely

$$ag^{(iv)} + 4bg'g''' + 6c(g')^2 g'' + 3b(g'')^2 + d(g')^4 = 0 \tag{5.4}$$

where

$$a = \gamma_2^3 + \gamma_1^2\gamma_2 + 2\frac{\gamma_1^4}{\gamma_2}; \quad b = \gamma_2^2 + 2\gamma_1^2 + \frac{\gamma_1^4}{\gamma_2^2}; \quad c = \gamma_2 + \frac{\gamma_1^4}{\gamma_2^3} + 2\frac{\gamma_1^2}{\gamma_2};$$

$$d = 1 + \frac{\gamma_1^4}{\gamma_2^4} + 2\frac{\gamma_1^2}{\gamma_2^2};$$

We observe that under the action of the spiral group, the original set of partial differential equations is not only uncoupled, but also transformed in such a way that the invariants $f(\eta)$ and $h(\eta)$ satisfy the same differential equation.

Consider equation (5.4) with $\gamma_1 = \gamma_2 = 1$ and $u(\eta) = g'(\eta)$. We then have

$$u''' + 4uu'' + 6u^2u' + 3(u')^2 + u^4 = 0. \tag{5.5}$$

This equation is linearized by raising the order. The transformation

$$y' = uy$$

gives rise to the sequence

$$y'' = y[u' + u^2]$$

$$y''' = y[u'' + 3uu' + u^3]$$

$$y^{(iv)} = y[u''' + 4uu'' + 6u^2u' + 3(u')^2 + u^4].$$

Consequently, the truth of (5.5) implies that

$$\frac{d^4y}{d\eta^4} = 0,$$

whereupon the solution for u is

$$\begin{aligned} u = y'/y &= \frac{3A_3\eta^2 + 2A_2\eta + A_1}{A_3\eta^3 + A_2\eta^2 + A_1\eta + A_0} \\ &= \frac{3\eta^2 + 2A\eta + B}{\eta^3 + A\eta^2 + B\eta + c}. \end{aligned} \tag{5.6}$$

Whereupon,

$$g(\eta) = \ln[\eta^3 + A\eta^2 + B\eta + C] + D. \tag{5.7}$$

Since both f and h must satisfy (5.4), it follows that $f(\eta) = h(\eta) = g(\eta)$ is an exact solution with four arbitrary constants A, B, C and D. Hence

$$F = w = G[\eta^3 + A\eta^2 + B\eta + C], \quad G = e^D, \quad \eta = x-y$$

constitute a set of exact solutions to the von Karman equations.

Remark 2.2. For the time independent problem other linear combinations of X_2, X_3, X_4, X_5 (without $\partial/\partial t$), X_6, X_8, X_9, X_{12}, X_{13}, X_{14} (with $f_i(t)$ = constants) can be used to generate solutions. Some of these may be the same as can be ascertained by examination of the commutator table and the splitting-nonsplitting analysis. That is not shown here.

6. THE TIME DEPENDENT EQUATIONS. Equations (1.4) and (1.5b) will be studied here using the dilatation generator X_5. The group invariants are

$$\zeta = x/\surd t; \quad \eta = y/\surd t; \quad F = f(\zeta,\eta); \quad w = h(\zeta,\eta). \tag{6.2}$$

This transformation results in the following partial differential equations for the functions f and h:

$$\Delta^2 f = (\frac{\partial^2 h}{\partial\zeta\partial\eta})^2 - \frac{\partial^2 h}{\partial\zeta^2}\frac{\partial^2 h}{\partial\eta^2}; \tag{6.3}$$

$$\begin{aligned}\Delta^2 h = &- \frac{1}{4}(\zeta^2 \frac{\partial^2 h}{\partial\zeta^2} + \eta^2 \frac{\partial^2 h}{\partial\eta^2} + 2\zeta\eta \frac{\partial^2 h}{\partial\zeta\partial\eta} + 3\zeta \frac{\partial h}{\partial\zeta} + 3\eta \frac{\partial h}{\partial\eta}) \\ &+ (\frac{\partial^2 f}{\partial\eta^2}\frac{\partial^2 h}{\partial\zeta^2} + \frac{\partial^2 f}{\partial\zeta^2}\frac{\partial^2 h}{\partial\eta^2} - 2\frac{\partial^2 f}{\partial\zeta\partial\eta}\frac{\partial^2 h}{\partial\zeta\partial\eta})\end{aligned} \tag{6.4}$$

If, at this point, we attempt to reduce equations (6.3) and (6.4) by re-applying the dilatation group, we discover that such a reduction is not possible. However, we can find solutions of this system by considering the action of the rotation group

$$\bar{\zeta} = \zeta\cos a - \eta\sin a;$$

$$\bar{\eta} = \zeta\sin a + \eta\cos a.$$

As indicated previously, the invariants of this transformation group are

$$z = \zeta^2 + \eta^2; \quad f = v(z); \quad h = u(z). \tag{6.5}$$

Substitution of these relations into (6.3) and (6.4) leads to two coupled ordinary differential equations for u and v, namely

$$\frac{d^2}{dz^2}\left[z^2\frac{d^2v}{dz^2}\right] = -\frac{1}{4}\frac{d}{dz}\left[z\left(\frac{du}{dz}\right)^2\right]; \tag{6.6}$$

$$\frac{d^2}{dz^2}\left[z^2\frac{d^2u}{dz^2}\right] = \frac{1}{2}\frac{d}{dz}\left[z\frac{du}{dz}\frac{dv}{dz}\right] - \frac{1}{16}\left[2z\frac{du}{dz} + z^2\frac{d^2u}{dz^2}\right]. \tag{6.7}$$

One set of solutions to this system is $u(z) = c$ and $v(z) = a_1 z\ln z + a_2\ln z + a_3 z + a_4$ where c and the a_i $(i = 1,\ldots,4)$ are constants. It thus follows that the time dependent von Karman equations possess the exact solution

$$F = a_1\left(\frac{x^2+y^2}{t}\right)\ln\left(\frac{x^2+y^2}{t}\right) + a_2\ln\left(\frac{x^2+y^2}{t}\right) + a_3\left(\frac{x^2+y^2}{t}\right) + a_4;$$

$$w = c$$

If we choose $u(z) = v(z)$ in equations (6.6) and (6.7), then we obtain the following equation for $v(z)$

$$\frac{3}{4}\frac{d}{dz}\left[z\left(\frac{dv}{dz}\right)^2\right] = \frac{1}{16}\left[2z\frac{dv}{dz} + z^2\frac{d^2v}{dz^2}\right],$$

from which a first integral can be obtained. It follows that

$$\frac{3}{4}z\left(\frac{dv}{dz}\right)^2 = \frac{1}{16}z^2\frac{dv}{dz} + \gamma, \tag{6.8}$$

where γ is a constant of integration. For $z \neq 0$ we can solve for dv/dz in (6.8) and obtain $v(z)$ (and thus $u(z)$) after a quadrature.

An alternative way of dealing with equations (6.3) and (6.4) is to look for solutions of the form $f = g(z)$ and $h = p(z)$ where $z = \zeta - \eta$. Under such a transformation, equations (6.3) and (6.4) become linear equations

$$\frac{d^4g}{dz^4} = 0;$$

$$\frac{d^4p}{dz^4} + \frac{1}{16}z^2\frac{d^2p}{dz^2} + \frac{3}{16}z\frac{dp}{dz} = 0. \tag{6.9}$$

The first of these equations can be readily integrated to give $g(z) = c_1 z + c_2 z + c_3 z + c_4$ for constants c_i $(i = 1,\ldots,4)$. Information about the stress distribution in the middle surface of the plate can then be obtained since

$$F = g(\frac{x-y}{\sqrt{t}})$$

implies that

$$\tau_{xy} = \sigma_x = \sigma_y = \frac{1}{t}\,[\frac{6c_1(x-y)}{\sqrt{t}} + 2c_2].$$

Equation (6.9) has two classical solutions $p_1' = \sin(z^2/8)$ and $p_2' = \cos(z^2/8)$ and a third solution of the form $p_3' = \int \sin(t-\tau)\tau^{-3/2}d\tau$, $t = z^2/8$. (These solutions are due to our colleagues Frank Stallard and Thomas Morley.)

7.0 OTHER POSSIBILITIES. The four groups considered (dilatation, rotation, translation and spiral) are not the only ones under which the von Karman equations remain invariant. To illustrate another (among the many) possibility consider $X_4 + \varepsilon X_1$, i.e., the invariants are obtained by integrating

$$[y \frac{\partial}{\partial x} - x \frac{\partial}{\partial y} + \varepsilon \frac{\partial}{\partial t}]I(x,y,t) = 0 \qquad (7.1)$$

where ε is a constant. The invariants are

$$r = (x^2+y^2)^{1/2} \quad \text{and} \quad I = \sin^{-1}\frac{x}{r} - \varepsilon t.$$

Consequently, solutions of the form

$$F = g[(x^2+y^2)^{1/2} \quad , \quad \sin^{-1}\frac{x}{r} - \varepsilon t]$$

with a corresponding form for w can be sought.

It should also be noted that these analyses yield invariant solutions of the differential equations alone. It may not be possible to match these exact invariant solutions to general auxiliary conditions such as prescribed boundary values or energy conservation integrals. However, these solutions are interesting since they are often asymptotic limits of other solutions.

REFERENCES

1. W. F. Ames, Nonlinear Partial Differential Equations in Engineering, Vol. II, Chapter 2, Academic Press, New York, 1972

2. G. W. Bluman and J. D. Cole, Similarity Methods for Differential Equations, Springer, New York, 1974.

3. L. V. Ovsiannikov, Group Analysis of Differential Equations (Russian Edition, NAUKA 1978); English translation edited by W. F. Ames, Academic Press, 1982.

4. R. E. Boisvert, Group Analysis of the Navier-Stokes Equations, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA 30332, 1982.

5. R. E. Boisvert, W. F. Ames and U. N. Srivastava, Group Properties and New Solutions of Navier-Stokes Equations, in press Jour. Engineering Mathematics.

6. F. Schwarz, Personal communication.

7. F. Schwarz, A Reduce Package for Determining Lie Symmetries of Ordinary and Partial Differential Equations, Comp. Physics Commun. 27, p. 179-186, 1982.

8. P. Roseneau and J. L. Schwarzmeier, Similarity Solutions of Systems of Partial Differential Equations Using MACSYMA, Courant Inst. of Math. Sci. Report No. COO-3077-160/MF-94, 1979.

9. K. A. Ames and W. F. Ames, On Group Analysis of the von Karman Equations, Int. Jl. Nonlinear Analysis: Theory, Methods and Applications, 6, 845, 1982.

10. S. P. Timoshenko and S. Woinowsky-Krieger, Theory of Plates and Shells, McGraw-Hill, New York, 1959.

11. J. S. W. Wong, On the Generalized Emden-Fowler Equations, SIAM Rev. 17, 339, 1975.

ON THE CONTROL OF A LINEAR STOCHASTIC SYSTEM WITH FINITE HORIZON[*]

P. L. Chow and J. L. Menaldi
Department of Mathematics, Wayne State University
Detroit, Michigan 48202

ABSTRACT. We consider a dynamic system whose state is governed by a linear stochastic differential equation with time-dependent coefficients. The control acts additively on the state of the system. Our objective is to minimize an integral cost which depends upon the evolution of the state and the total variation of the control process. It is proved that the optimal cost is the unique solution of an appropriate free boundary problem in a space-time domain. By using some decomposition arguments, the problems of a two-sided control, i.e. optimal corrections, and the case with constraints on the resources, i.e. finite fuel, can be reduced to a simpler case of only one-sided control, i.e. a monotone follower. These results are applied to solving some examples by the so-called method of similarity solutions.

I. INTRODUCTION. The optimal control of a stochastic system has a wide range of applications, such as optimal production scheduling, inventory control, investment policy, as well as the traditional problems in guidance and the

[*] This research has been supported in part by the U.S. Army Research Office, Contract DAAG29-83-K-0014.

trajectory tracking, (see, e.g. [1] - [9]). The stochastic models are commonly used in favor of the determinisitc ones because of the ever presence of noises and uncertainty.

In this article, we consider the evolution of a randomly perturbed system whose state, under an additive control action, is governed by an Itô's equation. The objective is to minimize the expected total cost over a finite horizon, with or without the constraint on the total resources. The methods of analysis are based on the dynamic programming technique [10] and the variational inequalities [11]. The main results to be presented here are taken from our recent joint paper [12]. We show that the optimal cost function may be characterized as a unique solution of a free-boundary problem for a certain parabolic equation. The optimal control is given by a reflected diffusion from the free boundary. Our decomposition and reduction theorems allow us to deal only with the one-sided control without resource constraint. This simplifies the general problem considerably. As a concrete example, a special case is solved explicitly with the aid of the so-called similarity solution. Even though the analysis is done in one dimension, most results may be extended to higher dimensions.

II. STATEMENT OF THE PROBLEM. In this paper, we wish to control a linear stochastic differential equation in the sense of Itô by using additive strategies, i.e. the evolution of the state is governed by

$$\left|\begin{array}{l} dy(s) = d\nu(s-t) + (a(s)y(s) + b(s))ds + \sigma(s)ds(s-t), \quad s \geq t \quad , \\ y(t) = x + \nu(0) \ , \end{array}\right. \qquad (1$$

where $a(s)$, $b(s)$ and $\sigma^2(s)$ stand for the drift and the covariance terms, and x is the initial state at the time t . $(\nu(s), s \geq 0)$ stands for the control which is a progressively measurable process with locally bounded variation.

To each control ν in $\mathcal{V}$, the class of admissible controls, it is associated with a cost given by the payoff functional

$$\left| \begin{aligned} J_{xt}(\nu) = E\{\int_t^T & f(y(s),s)\exp(-\int_t^s \alpha(\lambda)d\lambda)ds + c(t)\nu(0) \\ & + \int_t^T c(s)\exp(-\int_t^s \alpha(\lambda)d\lambda)d\nu(s-t)\} \end{aligned} \right. , \tag{2}$$

where f,α,c and T are respectively, the running cost, the discount factor, the instantaneous cost per unit of fuel and the finite horizon.

Our purpose is to characterize the optimal cost

$$\hat{u}(x,t) = \inf\{J_{xt}(\nu) : \nu \text{ in } \mathcal{V}\} \tag{3}$$

and to construct an optimal control $\hat{\nu}$, i.e.

$$\hat{\nu} \text{ in } \mathcal{V} \text{ such that } \hat{u}(x,t) = J_{xt}(\hat{\nu}) \tag{4}$$

for each initial state (x,t).

A similar study will be made for the optimal cost

$$\hat{v}(x,z,t) = \inf\{J_{xt}(\nu) : \nu \text{ in } \mathcal{V}, \nu(T) \le z\} , \tag{5}$$

where the positive constant z stands for the total amount of fuel available. This is associated with the previous cases under constraint of resources.

III. MAIN RESULTS. Consider the differential operators

$$Au = \frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2(t)\frac{\partial^2 u}{\partial x^2} - (a(t)x + b(t))\frac{\partial u}{\partial x} + \alpha(t)u \tag{6}$$

and

$$Bu = -\frac{\partial u}{\partial x} - c(t) \quad . \tag{7}$$

A heuristic application of the dynamic programming to the problem (1), (2) yields the following Hamilton-Jacobi-Bellman equation

$$\left| \begin{array}{l} (Au - f) \vee Bu = 0 \quad \text{in} \quad \mathbb{R} \times [0,T) \ , \\ u(\cdot,T) = 0 \quad \text{in} \quad \mathbb{R} \ , \end{array} \right. \tag{8}$$

where $x \vee y$ denotes the maximum of the two real numbers x and y. Equation (8) may be used to characterize the optimal cost $\hat{u}$ given by (4), for which the control is one-sided.

For the corresponding problem with finite resources, we have the following Hamilton-Jacobi-Bellman equation

$$\left| \begin{array}{l} (Av - f) \vee B'v = 0 \quad \text{in} \quad \mathbb{R} \times (0,\infty) \times [0,T) \ , \\ v(\cdot,\cdot,T) = 0 \quad \text{in} \quad \mathbb{R} \times [0,\infty) \ , \\ Av = f \quad \text{in} \quad \mathbb{R} \times \{0\} \times [0,T) \end{array} \right. \tag{9}$$

to be satisfied by the optimal cost $\hat{v}$ given by (5), where

$$B'v = \frac{\partial v}{\partial z} - \frac{\partial v}{\partial x} - c(t) \ . \tag{10}$$

We assume that $a(t)$, $b(t)$, $\sigma(t)$, $\alpha(t)$, $c(t)$ are Lipschitz functions from $[0,T]$ into $\mathbb{R}$ and either $c(t) \geq c_0 > 0$ for every t or $c(t) = 0$ for every t, and $f(x,t)$ is a smooth, positive function satisfying certain growth properties. Then the optimal cost $\hat{u}$ can be shown to be the unique solution of the following free boundary problem

$$\left| \begin{array}{l} A\hat{u} = f \quad \text{and} \quad B\hat{u} \leq 0 \quad \text{if} \quad x \geq x^*(t) \;, \\ A\hat{u} \leq f \quad \text{and} \quad B\hat{u} = 0 \quad \text{if} \quad x \leq x^*(t) \;. \end{array} \right. \tag{11}$$

where the free boundary $x = x^*(t)$, $t \in [0,T]$, is defined by

$$x^*(t) = \inf\{x : \frac{\partial \hat{u}}{\partial x} (x,t) + c(t) > 0\} \;. \tag{12}$$

In general we have obtained the following results:

(R1) <u>Decomposition Theorem</u>. Suppose that $\hat{u}(x,t)$ and $\hat{v}(x,z,t)$ are the optimal costs for the unlimited and limited resources, respectively, and $u^0(x,t)$ be the cost (2) without control $(\nu \equiv 0)$. Then the optimal cost $\hat{v}$ can be decomposed as follows:

$$\hat{v}(x,z,t) = \hat{u}(x,t) + h(x+z,t) \;, \tag{13}$$

where

$$h(x,t) = u^0(x,t) - \hat{u}(x,t) \;. \tag{14}$$

(R2) <u>Reduction Theorem</u>. If the running cost function f is symmetric about the trajectory $x = x_0$ so that

$$f(x,t) = f[2x_0(t) - x,t] + \beta(t)$$

for some function β , then the optimal cost $\hat{u}$ for a two-sided control has the property

$$\hat{u}(x,t) = \hat{u}(2x_0(t) - x,t) + \tilde{\beta}(t), (x,t) \quad \text{in} \quad \mathbb{R} \times [0,T] \;, \tag{15}$$

with

$$\tilde{\beta}(t) = \int_t^T [\beta(s) + b(s)c(s) + \dot{x}_0(s)c(s)]\exp(-\int_t^s \alpha(\lambda)d\lambda)ds . \tag{16}$$

Because of this symmetry, one may reduce the problem to the case of one-sided control.

(R3) Optimal Policy. Let $x = x^*(t)$ be the free boundary for a one-sided control. Then the optimal policy is to keep the evolution $y(t) \geq x^*(t)$ with the minimum use of resources, that is

$$\left| \begin{array}{l} y(s) \geq x^*(s) \quad \text{for} \quad s \geq t , \\ d\hat{\nu}(s-t) = 0 \quad \text{if} \quad y(s) > x^*(s) , \\ \nu(0) = \begin{cases} 0 , & \text{if} \quad x > x^*(t) \\ x^*(t) - x, & \text{otherwise} . \end{cases} \end{array} \right. \tag{17}$$

The optimal control is determined by the equation (1).

In view of the above results (R1) - (R3), to construct an optimal policy for the two-sided control with or without the resource limitation, we only need to solve the problem of the one-sided control with unlimited resources. Therefore our results yield a drastic simplification in treating the general problem.

IV. EXAMPLES. To illustrate our results, we shall consider some examples. We assume that the coefficients a,b,α,σ in (1) and (2) are constant, and the running cost $f(x)$ is time-independent. In addition, let $c(t) \equiv 0$, i.e., the cost for control is negligible. For $a > 0$ and $b < 0$, the equation (1) may be interpreted as an inventory model for which the rate of demand decreases as the stock increases. On the other hand, if $a < 0$ and $b > 0$, it becomes the Lagenvin equation describing the motion of a Brownian particle in the gravitational field. Then the problem pertains to the control of an Ornstein-Uhlenbeck process.

For unlimited resources, the average cost (2) yields

$$J_{xt}(\nu) = \{E\int_t^T f(y(s))e^{-\alpha s}ds\} \ . \tag{18}$$

By our result (R1), the optimal cost $\hat{u}$ must satisfy

$$\left| \begin{array}{l} A_0\hat{u} = f \quad \text{and} \quad \frac{\partial\hat{u}}{\partial x} \geq 0 \quad \text{if} \quad x \geq x^*(t) \ , \\ \\ A_0\hat{u} \leq f \quad \text{and} \quad \frac{\partial\hat{u}}{\partial x} = 0 \quad \text{if} \quad x \leq x^*(t) \ , \quad 0 \leq t \leq T \ . \end{array} \right. \tag{19}$$

where

$$A_0 u = -\frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2 \frac{\partial^2 u}{\partial x^2} - (ax+b)\frac{\partial u}{\partial x} + \alpha u \ , \tag{20}$$

$$x^*(t) = \inf\{x : \frac{\partial\hat{u}}{\partial x}(x,t) > 0\} \ . \tag{21}$$

To construct the solution $\hat{u}$ for $x \geq x^*(t)$, we let $s = (T-t)$ so that (11) gives the following free-boundary problem

$$\left| \begin{array}{l} v(x,s) \equiv \hat{u}(x,T-s) \ , \\ \\ Lv = \frac{\partial v}{\partial s} - \frac{1}{2}\sigma^2 \frac{\partial^2 v}{\partial x^2} - (ax+b)\frac{\partial v}{\partial x} + \alpha v = f(x) \ , \\ \\ \qquad \text{and} \\ \\ \frac{\partial v}{\partial x} \geq 0 \ , \quad \text{for} \quad x > x^*(T-s), \quad 0 \leq s \leq T \ , \\ \\ v(x,0) = 0 \ , \\ \\ \frac{\partial v}{\partial x}\Big|_{x = x^*(T-s)} = 0 \ , \end{array} \right. \tag{22}$$

where $v(x,s) = \hat{u}(x,T-s)$.

By a proper change of variables, the system (22) may be reduced to a free-boundary problem for a heat equation. Then, by applying the method of similarity solution, the free-boundary is found to be

$$x^*(t) = \frac{\delta}{2}\left[\frac{1-e^{-2a(T-t)}}{a}\right]^{1/2} - \frac{b}{a}\,, \quad 0 \leq t \leq T\,, \tag{23}$$

where the parameter δ is determined by the equation

$$\delta^2 + 1 = \frac{4\delta\varphi_1(\delta)\int_\eta^\infty [\varphi_1(\lambda)]^{-2} e^{-1/2(\lambda^2) - \delta^2)} d\lambda}{\varphi_1(\delta)\varphi_1'(\delta)\int_\eta^\infty [\varphi_1(\lambda)]^{-2} e^{-1/2\lambda^2} d\lambda - 1}\,. \tag{24}$$

This may be solved numerically to yield $\delta = -0.6388\ldots$.

Now, suppose the resource ν for control is finite so that $0 \leq \nu(T) \leq z$. The optimal cost $\hat{v}(x,z,t)$ can be decomposed, according to (R1) into two simple problems. That is,

$$\hat{v}(x,z,t) = u^0(x+z,t) - [\hat{u}(x+z,t) - \hat{u}(x,t)] \tag{25}$$

where $\hat{u}(x,t)$ is the optimal cost without resource constraint, while $u^0(x,t)$ is the cost of free evolution. Therefore it must satisfy

$$\left|\begin{array}{l} A_0 u^0 = f\,, \quad 0 \leq t < T\,, \quad x \in \mathbb{R}\,, \\ u^0(x,T) = 0\,, \\ u^0(x,t) = O(|f(x)|) \quad \text{as} \quad |x| \to \infty\,, \end{array}\right. \tag{26}$$

where A_0 is defined by (20). The equation (26) may be solved to give

$$\left|\begin{aligned} u^0(x,t) &= e^{-\alpha(T-t)}\int_0^{(T-t)}\int_{\mathbb{R}} \frac{\exp(\frac{[\xi(x,t)-\rho]^2}{2[\tau(t)-\lambda]} + 2a\beta\lambda)}{2\pi[\tau(t)-\lambda]} \\ &\qquad \times\ (1+2a\lambda)^{\beta} f[\sigma(1+2a\lambda)^{-1/2}\ \rho - \frac{b}{a}]d\lambda d\rho\ , \\ \xi(x,t) &= \frac{1}{\sigma}(x+\frac{b}{a})\ e^{a(T-t)}\ , \\ \tau(t) &= (2a)^{-1}[e^{2a(T-t)} - 1]\ . \end{aligned}\right. \tag{27}$$

Note that the free boundary, given by (23), remains unchanged. In particular, for $m = 2$, this problem may be solved explicitly.

We wish to point out that, for the optimal correction problems, the case of vanishing cost, $c = 0$, is less interesting. In this case the optimal policy would be to counteract the noise as long as the resources remain available so that $f(y(t),t)$ is kept to the minimum.

References

[1] J.A. Bather, A Continuous Time Inventory Model, *J. Appl. Prob.*, *3* (1966), pp. 538-549.

[2] J.A. Bather, A Diffusion Model for the Control of a Dam, *J. Appl. Prob.*, *5* (1968), pp. 55-71.

[3] J.A. Bather and H. Chernoff, Sequential Decisions in the Control of a Spaceship, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1967, Vol.3, pp. 181-207.

[4] J.A. Bather and H. Chernoff, Sequential Decisions in the Control of a Spaceship (Finite Fuel), *J. Appl. Prob.*, *4* (1967), pp. 584-604.

[5] V.E. Benes, L.A. Shepp and H.S. Witsenhausen, Some Solvable Stochastic Control Problems, *Stochastics*, *4* (1980), pp. 39-83.

[6] I. Karatzas, The Monotone Follower Problem in Stochastic Decision Theory, *Appl. Math. Optim.*, *7* (1981), pp. 175-189.

[7] J.L. Menaldi, J.P. Quadrat and E. Rofman, On the Role of the Impulse Fixed Cost in Stochastic Optimal Control: An Application to the Management of Energy Production, *Proceedings of the Tenth IFIP Conference on System Modelling and Optimization*, Lecture Notes in Control and Information Sciences, *38* (1982), Springer-Verlag, New York, pp. 671-679.

[8] J.L. Menaldi and M. Robin, On Some Cheap Control Problems for Diffusion Processes, *Trans. Am. Math. Soc.*, (1983), to appear. See also *C.R. Acad. Sc. Paris, Serie* I, 294 (1982), pp. 541-544.

[9] J.L. Menaldi and E. Rofman. A Continuous Multi-Echelon Inventory Problem, *Proceedings of the Fourth IFAC-IFIP Symposium on Information Control Problems in Manufacturing Techonology*, Gaithersburg, Maryland, USA, October 1982, pp. 41-49.

[10] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[11] A. Bensoussan and J.L. Lions, *Applications des inequations variationnelles en controle stochastique*, Dunod, Paris, 1978.

[12] P.L. Chow, J.L. Menaldi and M. Robin, Additive Control of Stochastic Linear Systems with Finite Horizon, submitted to SIAM J. Control Optim.

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS PESEARCH CENTER

A NONLINEAR INTEGRAL EQUATION OCCURRING IN A SINGULAR FREE BOUNDARY PROBLEM

Klaus Höllig[1,2] and John A. Nohel[1]

ABSTRACT

We study the Cauchy problem

$$\begin{cases} u_t = \phi(u_x)_x, & (x,t) \in \mathbf{R} \times \mathbf{R}_+ , \\ u(\cdot,0) = f & \end{cases}$$

with the piecewise linear constitutive function $\phi(\xi) = \xi_+ = \max(0,\xi)$ and with smooth initial data f which satisfy $xf'(x) > 0$, $x \in \mathbf{R}$, and $f''(0) > 0$. We prove that the free boundary s, given by $u_x(s(t)^+,t) = 0$, is of the form

$$s(t) = -\kappa\sqrt{t} + o(\sqrt{t}), \quad t \to 0^+ ,$$

where the constant $\kappa = 0.9037...$ is the (numerical) solution of a particular nonlinear equation. Moreover, we show that for any $\alpha \in (0,1/2)$,

$$|\frac{d^2}{dt^2} f(s(t))| = O(t^{\alpha-1}), \quad t \to 0^+ .$$

The proof involves the analysis of a nonlinear singular integral equation.

AMS (MOS) Subject Classifications: 35K55, 35K65, 45G05

Key Words: Cauchy problem, parabolic, nonlinear, free boundary regularity, nonlinear singular integral equation

[1] Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

[2] This material is based upon work partially supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2.

A NONLINEAR INTEGRAL EQUATION OCCURRING IN A SINGULAR FREE BOUNDARY PROBLEM

Klaus Höllig[1,2] and John A. Nohel[1]

1. Introduction and Result. We study the Cauchy problem

$$(1) \qquad \begin{cases} u_t = \phi(u_x)_x, & (x,t) \in \mathbb{R} \times \mathbb{R}_+ , \\ u(\cdot,0) = f \end{cases}$$

with the piecewise linear constitutive function $\phi : \mathbb{R} \to \mathbb{R}_+$ given by $\phi(\xi) = \xi_+ = \max(\xi,0)$; the initial data $f : \mathbb{R} \to \mathbb{R}$ are assumed smooth, specifically $f \in C^3(\mathbb{R})$ with bounded derivatives, and satisfy the conditions

$$(2) \qquad \begin{cases} xf'(x) > 0, & x \in \mathbb{R} , \\ f''(0) > 0 . \end{cases}$$

One motivation for the study of the Cauchy problem (1), (2) is its similarity with the well-known one phase Stefan problem (in one space dimension) [3,4,7,8] in which one would assume $f'(x) \equiv -1$ for $x < 0$, as well as $f'(x) > 0$ for $x > 0$, so that f' has a jump discontinuity at $x = 0$. The assumption (2) yields a different behavior of the solution u and of the resulting free boundary. Indeed, here (c.f. the Theorem below), the free boundary s, given by $u_x(s(t)^+,t) = 0$, is of the form

$$(3) \qquad s(t) = -\kappa\sqrt{t} + O(t^{1/2+\alpha}), \quad t \to 0^+ ,$$

where κ is a positive constant and $0 < \alpha < 1/2$. Thus, the function s is not (infinitely) differentiable at $t = 0$, contrary to the situation for the Stefan problem [7].

The result (3) is established by solving a nonlinear integral equation ((15) below) with kernels which depend on the unknown function s and which are also singular in the

[1] Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

[2] This material is based upon work partially supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2.

sense that the integral on $(0,t)$ of the kernel does not approach zero as $t \to 0^+$. One consequence of this is that the integral operator defined by (15) is not compact in a suitable Hölder class.

The principal motivation for the study of the Cauchy problem (1), (2) is that it serves as a prototype of nonlinear parabolic problems which arise as monotone "convexifications" of nonlinear diffusion equations with nonmonotone constitutive functions ϕ (see [5] and [6]); in [6, section 4] the reader will also find the formulation and preliminary analysis of such a convexified problem, corresponding to a piecewise linear nonmonotone ϕ (specificallly, $\phi'((-\infty,a) \cup (b,\infty)) > 0$, $\phi'(a,b) < 0$, $0 < a < b < \infty$). The analysis in [5] shows the existence of infinitely many solutions u of the nonmonotone problem, each having u_x bounded, and u_x omitting the values in $[a,b]$; thus each solution u exhibits phase changes. Numerical experiments further suggest the conjecture that the "physically correct" solution of the nonmonotone problem is the one which, as $t \to \infty$, approaches the unique solution of the appropriately related convexified monotone problem. However, for small $t > 0$ the behavior of the solution of (1), (2) is qualitatively different (see (3)). The present study of (1), (2) is intended as a step towards the understanding of this intriguing phenomenon. The relation of the convexified problem in [6] to the Cauchy problem (1), (2) is clear (the particular boundary conditions in [6] do not play a role in the analysis of the free boundary curve).

It is simple to give a formal explanation for (3). We rewrite (1), (2) as the free boundary problem

$$\text{(4a)} \qquad \begin{cases} u_t = u_{xx}, & s(t) < x < \infty, \quad t \in \mathbb{R}_+ , \\ u_x(s(t),t) = 0 & \\ u(\cdot,0) = f . & \end{cases}$$

From the constitutive function ϕ one also has the equation

$$\text{(4b)} \qquad \begin{cases} u_t = 0, & -\infty < x < s(t), \quad t \in \mathbb{R}_+ \\ u(\cdot,0) = f . & \end{cases}$$

Therefore, assuming the continuity of u across the free boundary $s(t)$ and assuming that s is monotone decreasing (c.f. paragraph preceding the Theorem), we have

$$(5) \qquad \begin{cases} u(s(t),t) = f(s(t)), & t \in \mathbb{R}_+ , \\ s(0) = 0 . \end{cases}$$

Differentiating (5) with respect to t and using $u_x(s(t)^+,t) = 0$, where "+" denotes the limit from the right, we obtain

$$(6) \qquad f'(s(t))s'(t) = u_{xx}(s(t)^+,t) .$$

Since by the assumption (2)

$$f'(x) = f''(0)x + O(|x|^2), \quad |x| \to 0^+ ,$$

a simple calculation formally yields (3) with $\kappa = \sqrt{2}$ (provided one assumes continuity from the right of u_t and u_{xx} up to the free boundary s).

The rigorous treatment of the problem consists of analyzing in Section 3 the nonlinear integral equation (15) for the free boundary $x = s(t)$. Our analysis shows that (3) holds, but that the constant κ is the solution of the nonlinear equation (16); its numerical value is $\kappa = 0.9037\ldots$, and not $\kappa = \sqrt{2}$ which was predicted by the above formal calculation. It also follows that $s(t)$ is smooth for $t > 0$ thus justifying (5) and (6) for positive t; in particular one sees from (6) that s is as smooth as the initial function f is. We remark that for $t > \varepsilon > 0$ the problem (1), (2) can also be viewed as a one phase Stefan problem; consequently the results in Kinderlehrer and Nirenberg [7] yield the regularity of the free boundary for $t > 0$.

The existence of a unique generalized continuous solution for problem (1), and hence of a unique free boundary, follows from nonlinear semigroup theory for m-accretive operators [1,2]. Approximating (1) by the implicit Euler scheme one can also show the existence of the free boundary s which is Hölder continuous on $[0,\infty)$ with exponent $1/2$ and monotone decreasing. However, using such general methods, it is not possible to analyze the precise behavior of s at $t = 0$.

Our main result is:

THEOREM. *Define*

$$(7) \qquad r(t) = \frac{d}{dt} f(s(t)) .$$

Then for any $\alpha \in (0,1/2)$ *there exists* $T > 0$ *such that* r *is continuous on* $[0,T]$ *and satisfies*

(8) $$t^{1-\alpha}|r'(t)| \le c(f), \quad 0 < t \le T ,$$

where $c(f) > 0$ *is a constant which depends on the data* f. *Moreover,* (3) *holds with*

$$\kappa := \left(2\,\frac{r(0)}{f''(0)}\right)^{1/2} = 0.9037\ldots \quad .$$

The constant κ *is the (numerical) solution of equation* (16) *in Section 3.*

By the definition of κ, the result (3) follows from (7) and the assertion (8). To see this, we solve (7) for s. Let $R(t) = \int_0^t r(\tau)d\tau$ and integrate (7) obtaining

$$R(t) = f(s(t)) - f(0) .$$

Define the function g implicitly by

$$g(-\text{sign}(x)\sqrt{f(x) - f(0)}) = x .$$

Since we assume that

(9) $$f(x) - f(0) = \beta^2 x^2 + O(|x|^3), \quad |x| \to 0^+ ,$$

$(\beta^2 = f''(0)/2)$, g is well defined for small $|x|$ and

(10) $$g(x) = -\beta^{-1}x + O(|x|^2), \quad |x| \to 0^+ .$$

For a small interval $[0,T]$, the monotone decreasing solution of (7) is given by

(11) $$s(t) = g(\sqrt{R(t)}), \quad 0 \le t \le T ,$$

and (3) follows from (8) and (10).

The Theorem describes the regularity of the free boundary at $t = 0$. It is sharp in the sense that, unless $f'''(0) = 0$, the estimate (8) does not hold for $\alpha > 1/2$ (c.f. the Remark at the end of the paper in Section 3).

It should also be observed that the second derivatives of the solution u are not continuous at the point $(x,t) = (0,0)$, because using (6), (7) and the definition of κ one has

$$\lim_{t\to 0^+} u_{xx}(s(t)^+,t) = r(0) = \frac{\kappa^2}{2} f''(0) \ne \lim_{x\to 0^+} u_{xx}(x,0) = f''(0) .$$

However, on the set $\{t : f'(s(t)) < 0\}$ the free boundary s is as smooth as the function f. This can be shown by a bootstrap argument, using standard regularity results for the heat equation on a domain with curved boundaries. We believe that the Theorem can be extended to a general monotone constitutive function ϕ with $\phi'(\cdot)$ discontinuous at 0 and with $\phi'(\xi) > c > 0$, $\xi \in \mathbb{R}_+$; the corresponding value of κ will depend on $\phi'(0^+)$.

The Theorem is proved in Section 3 by solving an integral equation for the function r derived in Section 2.

We are grateful for helpful discussions with our colleagues Tom Beale, Carl de Boor, Michael Crandall and Emmanuel DiBenedetto; we also thank Fred Sauer for the numerical computations.

2. The Integral Equation for the Free Boundary.

Let

$$\Gamma(x,t) := \frac{1}{2\sqrt{\pi}}\, t^{-1/2} \exp\left(-\frac{x^2}{4t}\right)$$

denote the fundamental solution of the heat equation. Let $v := u_x$ be the solution of the problem

$$\text{(4a')} \qquad \begin{cases} v_t = v_{xx}, \ (x,t) \in \Omega_T := \{(x,t) : x > s(t),\ t \in (0,T)\}\ , \\ v(s(t),t) = 0 \\ v(\cdot,0) = f' \end{cases}$$

and assume that the free boundary s satisfies $s \in C[0,T] \cap C^1(0,T]$. Integrating Green's identity

$$\frac{\partial}{\partial\xi}\left(\Gamma(x-\xi,t-\tau)v_\xi(\xi,\tau) - \frac{\partial}{\partial\xi}\Gamma(x-\xi,t-\tau)v(\xi,\tau)\right)$$

$$- \frac{\partial}{\partial\tau}\left(\Gamma(x-\xi,t-\tau)v(\xi,\tau)\right) = 0$$

over the domain Ω_t we obtain, for $x > s(t)$, the representations

$$\text{(12)} \qquad v(x,t) = \int_0^\infty \Gamma(x-\xi,t)f'(\xi)d\xi - \int_0^t \Gamma(x-s(\tau),t-\tau)v_\xi(s(\tau),\tau)d\tau\ ,$$

$$\text{(13)} \qquad v_x(x,t) = \int_0^\infty \Gamma(x-\xi,t)f''(\xi)d\xi - \int_0^t \Gamma_x(x-s(\tau),t-\tau)v_\xi(s(\tau),\tau)d\tau\ .$$

Passing to the limit $x \to s(t)^+$ in (13) yields

$$(14) \qquad r(t) = 2\int_0^\infty \Gamma(s(t) - \xi, t) f''(\xi)\,d\xi - 2\int_0^t \Gamma_x(s(t) - s(\tau), t - \tau) r(\tau)\,d\tau ,$$

where (see (6) and (7)) $r(t) = \frac{d}{dt} f(s(t)) = v_x(s(t),t)$. The justification for this passage to the limit is contained in the following result.

LEMMA 1. _If_ $s \in C([0,T]) \cap C^1((0,T])$ _and_ $r \in C([0,T])$, _we have for_ $t < T$

$$\lim_{x \searrow s(t)} \int_0^t [\Gamma_x(s(t) - s(\tau), t - \tau) - \Gamma_x(x - s(\tau), t - \tau)] r(\tau)\,d\tau = \frac{1}{2} r(t) .$$

Proof. We write

$$\int_0^t [\cdots] r\,d\tau =$$

$$\frac{1}{4\sqrt{\pi}} \int_0^t \frac{s(t) - s(\tau)}{(t - \tau)^{3/2}} \left[\exp\left(-\frac{(x - s(\tau))^2}{4(t - \tau)}\right) - \exp\left(-\frac{(s(t) - s(\tau))^2}{4(t - \tau)}\right)\right] r(\tau)\,d\tau$$

$$+ \frac{1}{4\sqrt{\pi}} \int_0^t \frac{x - s(t)}{(t - \tau)^{3/2}} \left[\exp\left(-\frac{(x - s(\tau))^2}{4(t - \tau)}\right) - \exp\left(-\frac{(x - s(t))^2}{4(t - \tau)}\right)\right] r(\tau)\,d\tau$$

$$+ \frac{1}{4\sqrt{\pi}} \int_0^t \frac{x - s(t)}{(t - \tau)^{3/2}} \exp\left(-\frac{(x - s(t))^2}{4(t - \tau)}\right) r(\tau)\,d\tau =: \sum_{\nu=1}^{3} \int_0^t I_\nu .$$

In view of the assumptions on s and r it is easy to see that, for $\nu = 1,2$,

$$\left|\int_0^t I_\nu\right| < \left|\int_{t-\delta}^t I_\nu\right| + \left|\int_0^{t-\delta} I_\nu\right| < O(\sqrt{\delta}) + c_\delta o(|x - s(t)|)$$

which implies that

$$\lim_{x \searrow s(t)} \int_0^t I_\nu = 0, \qquad \nu = 1,2 .$$

Finally, $\frac{1}{4\sqrt{\pi}} \int_{-\infty}^{t} \frac{y}{(t-\tau)^{3/2}} \exp\left(-\frac{y^2}{4(t-\tau)}\right) d\tau = \frac{1}{2}$ implies that

$$\lim_{x \searrow s(t)} \int_0^t I_3 = \frac{1}{2} r(t) \; .$$

3. *Proof of the Theorem.* We write the integral equation (14) in the form

$$r(t) = \frac{1}{\sqrt{\pi}} \int_0^\infty \exp\left(-\frac{1}{4}\left(\frac{s(t)}{\sqrt{t}} - \xi\right)^2\right) f''(\xi\sqrt{t})\, d\xi \tag{15}$$

$$+ \frac{1}{\sqrt{\pi}} \int_0^1 \frac{A(s,t,\tau)}{1-\tau} \exp(-A(s,t,\tau)^2) r(t\tau)\, d\tau =: (Fr)(t) + (Kr)(t) \; ,$$

where

$$A(s,t,\tau) := \frac{s(t) - s(t\tau)}{2(t - t\tau)^{1/2}} \; .$$

It will be convenient to introduce the class of functions $H^\alpha[0,T]$, $0 < \alpha < 1$, defined by

$$H^\alpha[0,T] = \{\rho : [0,T] \to \mathbf{R} : |\rho|_\alpha := \sup_{0<t\leq T} t^{1-\alpha} |\rho'(t)| < \infty\} \; .$$

The class H^α is obviously contained in the Hölder-class with exponent α.

The Theorem is a consequence of:

PROPOSITION. *For any* $\alpha \in (0,1/2)$, *the integral equation* (15), *with* s *related to* r *by* (11), *has a solution* $r \in H^\alpha[0,T]$ *for some* $T > 0$. *The constant* $\kappa := \sqrt{r(0)}/\beta$ ($\beta^2 = \frac{1}{2} f''(0)$) *does not depend on* f *and is implicitly determined by the equation*

$$\frac{4}{\sqrt{\pi}} \int_0^\infty \exp\left(-\left(\frac{\kappa}{2} + \xi\right)^2\right) d\xi = \tag{16}$$

$$\kappa^2\left(1 + \frac{1}{\sqrt{\pi}} \int_0^1 \frac{\kappa}{2} \frac{1}{\sqrt{1-\tau}(1+\sqrt{\tau})} \exp\left(-\frac{\kappa^2}{4} \frac{1-\sqrt{\tau}}{1+\sqrt{\tau}}\right) d\tau\right) \; ;$$

the numerical value of κ *is* 0.9037... .

REMARK. The Proposition does not assert uniqueness of the function r (hence of the free boundary s) which could be established by showing that the operator $F + K$ in (15) is a strict contraction; this is technically even more complicated than our proof. However, the uniqueness of r is a consequence of the uniqueness of solutions of the original problem (1) discussed in the Introduction.

We prove the Proposition by iterating the integral equation (15) in the form

$$r_{n+1} = Fr_n + Kr_n, \quad n \in \mathbb{N}, \tag{17}$$

with $r(0) = r_0 = \kappa^2\beta^2$, where κ is the solution of (16) and $\beta^2 = \frac{1}{2} f''(0)$.

We shall show as a consequence of Lemmas 2 and 3 below that, for $r \in H^\alpha$ with $r(0) = \kappa^2\beta^2$,

$$\lim_{t\to 0^+} (Fr)(t) = \frac{1}{\sqrt{\pi}} \int_0^\infty \exp\left(-\frac{1}{4}(\kappa + \xi)^2\right) 2\beta^2 d\xi, \tag{18}$$

$$\lim_{t\to 0^+} (Kr)(t) = -\frac{1}{\sqrt{\pi}} \int_0^1 \frac{\kappa}{2} \frac{1}{\sqrt{1-\tau}(1+\sqrt{\tau})} \exp\left(-\frac{\kappa^2}{4} \frac{1-\sqrt{\tau}}{1+\sqrt{\tau}}\right) \kappa^2\beta^2 d\tau. \tag{19}$$

Since κ is the solution of (16), this implies that $r_n(0) = \kappa^2\beta^2$ for $n \in \mathbb{N}$.

Moreover, we shall establish the a priori estimates: for $r \in H^\alpha[0,T]$, $0 < \alpha < 1/2$,

$$|Fr|_\alpha < c(T) + (c_1(\alpha) + c(T))|r|_\alpha, \tag{20}$$

where $c_1(\alpha) = \dfrac{\kappa^{-1}}{\sqrt{\pi}(1+\alpha)} \exp\left(-\frac{1}{4}\kappa^2\right)$, and

$$|Kr|_\alpha < (c_2(\alpha) + c(T))|r|_\alpha, \tag{21}$$

where $c_2(\alpha) = c_{21}(\alpha) + c_{22}(\alpha)$ with

$$c_{21}(\alpha) = \frac{\kappa}{2\sqrt{\pi}} \int_0^1 \frac{\tau^\alpha}{\sqrt{1-\tau}(1+\sqrt{\tau})} \exp\left(-\frac{\kappa^2}{4} \frac{1-\sqrt{\tau}}{1+\sqrt{\tau}}\right) d\tau,$$

$$c_{22}(\alpha) = \frac{\kappa\left(1 + \frac{1}{2+2\alpha}\right)}{\sqrt{\pi}(2+4\alpha)} \int_0^1 \frac{1-\tau^{1/2+\alpha}}{(1-\tau)^{3/2}} \left(1 - \kappa\frac{\sqrt{1-\tau}}{1+\sqrt{\tau}}\right) \times \exp\left(-\frac{\kappa^2}{4} \frac{\sqrt{1-\tau}}{1+\sqrt{\tau}}\right) d\tau,$$

and where $c(T)$ is a constant such that $c(T) \to 0$ as $T \to 0^+$, uniformly for

$r \in \{\rho : |\rho(0)| + |\rho|_\alpha < \text{const.}\}$.

We first use the estimates (20), (21) to complete the proof of the Proposition. Combining the estimates (20) and (21) one has

$$(22) \qquad |r_{n+1}|_\alpha < c(T) + (c_1(\alpha) + c_2(\alpha) + c(T))|r_n|_\alpha \ .$$

Crucial for the following argument is the fact that

$$c_1(\tfrac{1}{2}) + c_2(\tfrac{1}{2}) = 0.339\ldots + 0.453\ldots =: \omega < 1 \ .$$

Set $\bar{\omega} := \frac{1+\omega}{2} < 1$ and choose $\alpha \in (0,1/2)$ close to $1/2$ and $T > 0$ such that for all $r \in H^\alpha$ with $r(0) = \kappa^2\beta^2$ and $|r|_\alpha < \frac{1}{1-\bar{\omega}}$

$$c_1(\alpha) + c_2(\alpha) + c(T) < \bar{\omega} \ .$$

It should be observed that if one chooses $\alpha > 1/2$ then we cannot prove the crucial estimate (20), cf. e.g. (24). By (22), we have

$$|r_n|_\alpha < \frac{1}{1-\bar{\omega}}, \qquad n \in \mathbb{N} \ .$$

Hence we can select a subsequence of r_n which converges in $C[0,T]$ to a function $r_\infty \in H^\alpha[0,T]$ with $r_\infty(0) = \kappa^2\beta^2$. Set $s_n := g(\sqrt{R_n})$. To pass to the limit in (17) note that by Lemmas 2 and 3 below the expressions $\exp\left(-\frac{1}{4}\left(\frac{s_n(t)}{\sqrt{t}} - \xi\right)^2\right)$ and $\frac{A(s_n,t,\tau)}{1-\tau}\exp(-A(s_n,t,\tau)^2)$ converge pointwise (for $n \to \infty$) and are majorized by integrable functions, uniformly in $n \in \mathbb{N}$. This completes the proof of the Proposition and of the Theorem.

It remains to establish the assertions (18)-(21). We require two auxiliary results. We denote by c a generic constant which may depend on α, $|r|_\alpha$ and T, and we assume throughout that $T = T(|r|_\alpha, \alpha)$ is sufficiently small.

LEMMA 2. For $r \in H^\alpha$, $\alpha \in (0,1/2)$, with $r(0) = \kappa^2\beta^2$ we have

$$|s(t) + \kappa\sqrt{t}| < ct^{1/2+\alpha} \ .$$

Proof. Note that $|r(t) - r(0)| < ct^\alpha$ and therefore $|R(t) - r(0)t| < ct^{1+\alpha}$. Using (10), (11) and this inequality one has

$$|s(t) + \kappa\sqrt{t}| = |g(\sqrt{R(t)}) + \kappa\sqrt{t}| < |-\beta^{-1}\sqrt{R(t)} + \kappa\sqrt{t}| + ct < ct^{1/2+\alpha} + ct \ .$$

LEMMA 3. For $r \in H^{\alpha}$ with $r(0) = \kappa^2\beta^2$ we have

$$|A(s,t,\tau)| < c\,\frac{1-\sqrt{\tau}}{\sqrt{1-\tau}} = c\,\frac{\sqrt{1-\tau}}{1+\sqrt{\tau}}\,.$$

Proof. Using $f'(s)s' = r$, (9) and Lemma 2, we obtain

$$|s(t) - s(t\tau)| = \left|\int_{t\tau}^{t} \frac{r(\sigma)}{f'(s(\sigma))}\,d\sigma\right| < c\int_{t\tau}^{t} (2\beta^2\kappa\sqrt{\sigma} - c\sigma^{1/2+\alpha})^{-1}d\sigma < c(\sqrt{t} - \sqrt{t\tau})\ ;$$

this establishes the claim by the definition of $A(s,t,\tau)$.

Lemma 3 shows that the kernel corresponding to the operator K in (15) is integrable. Moreover, we see from Lemma 2 that

$$A_0(\tau) := \lim_{t\to 0^+} A(s,t,\tau) = -\frac{\kappa}{2}\,\frac{1-\sqrt{\tau}}{\sqrt{1-\tau}}\,. \tag{23}$$

Using this and Lemma 2, we can pass to the limit in (15), thus establishing (18) and (19).

Proof of (20). To estimate the norm of Fr, use the definition in (15) to form

$$\frac{d(Fr)(t)}{dt} = \frac{1}{\sqrt{\pi}}\int_0^{\infty} \exp\left(-\frac{1}{4}\left(\frac{s(t)}{\sqrt{t}} - \xi\right)^2\right)\frac{1}{2}\,t^{-1/2}\xi f'''(\xi\sqrt{t})d\xi$$

$$-\left[\frac{1}{\sqrt{\pi}}\int_0^{\infty}\frac{1}{2}\left(\frac{s(t)}{\sqrt{t}} - \xi\right)\exp\left(-\frac{1}{4}\left(\frac{s(t)}{\sqrt{t}} - \xi\right)^2\right)f''(\xi\sqrt{t})d\xi\right] \times \frac{d}{dt}\,\frac{s(t)}{\sqrt{t}}\,.$$

As $t \searrow 0$, the term in square brackets tends (use (9)) to

$$\frac{1}{\sqrt{\pi}}\int_0^{\infty}\frac{1}{2}(-\kappa - \xi)\exp\left(-\frac{1}{4}(-\kappa - \xi)^2\right)2\beta^2 d\xi$$

$$= -\frac{2}{\sqrt{\pi}}\,\beta^2\exp\left(-\frac{1}{4}\,\kappa^2\right) = 2(1+\alpha)\kappa\beta^2 c_1(\alpha)\,.$$

Therefore,

$$\left|\frac{d(Fr)(t)}{dt}\right| < ct^{-1/2} + (2(1+\alpha)\kappa\beta^2 c_1(\alpha) + \bar{c}(t))\left|\frac{d}{dt}\left(\frac{s(t)}{\sqrt{t}}\right)\right|\,. \tag{24}$$

It remains to estimate $\frac{d}{dt}\frac{s(t)}{\sqrt{t}}$. Using (7), Lemma 2, (9) and (10) we have

$$\left|\frac{s'(t)}{\sqrt{t}} - \frac{1}{2}\frac{s(t)}{t^{3/2}}\right| = t^{-3/2}\left|\frac{1}{f'(s(t))}\right|\left|tr(t) - \frac{1}{2}s(t)f'(s(t))\right|$$

$$< t^{-3/2}t^{-1/2}(\tfrac{1}{2}\beta^{-2}\kappa^{-1} + \bar{c}(t))[|tr(t) - \beta^2 s(t)^2| + \bar{c}t^{3/2}]$$

$$< t^{-1}(\tfrac{1}{2}\beta^{-2}\kappa^{-1} + \bar{c}(t))[|tr(t) - R(t)| + \bar{c}t^{3/2}] \ .$$

A simple calculation shows that

(25) $$|tr(t) - R(t)| < \frac{1}{1+\alpha}t^{1+\alpha}|r|_\alpha ,$$

and this yields

(26) $$\left|\frac{d}{dt}\left(\frac{s(t)}{\sqrt{t}}\right)\right| < t^{\alpha-1}(\tfrac{1}{2}\beta^{-2}\kappa^{-1} + \bar{c}(t))(\frac{1}{1+\alpha} + \bar{c}(t))|r|_\alpha \ .$$

Combining (24) and (26) proves (20).

We next turn to the proof of (21). We write (cf. (15))

$$\frac{d}{dt}(Kr)(t) = \frac{1}{\sqrt{\pi}}\int_0^1 \frac{1}{1-\tau} A \exp(-A^2)\tau r'(t\tau)d\tau$$

$$+ \frac{1}{\sqrt{\pi}}\int_0^1 \frac{1}{1-\tau}(1 - 2A^2)\exp(-A^2)\frac{dA}{dt}r(t\tau)d\tau =: (K_1r)(t) + (K_2r)(t)$$

and estimate each term separately.

(i) Since $|r'(t\tau)| < (t\tau)^{\alpha-1}|r|_\alpha$, it follows from (23) that

(27) $$|(K_1r)(t)| < (c_{21} + c(t))t^{\alpha-1}|r|_\alpha \ .$$

(ii) To estimate K_2r we first consider the term $\frac{d}{dt}A(s,t,\tau)$. Using the definition of A and (7), we obtain

$$2(t - t\tau)^{1/2}t\frac{d}{dt}\left(\frac{1}{2}\frac{s(t) - s(t\tau)}{(t - t\tau)^{1/2}}\right) = ts'(t) - (t\tau)s'(t\tau) - \frac{1}{2}s(t) + \frac{1}{2}s(t\tau) =$$

$$\int_{t\tau}^{t}\frac{d}{d\sigma}(\sigma s'(\sigma) - \frac{1}{2}s(\sigma))d\sigma = \int_{t\tau}^{t}\left(\frac{1}{2}s'(\sigma) + \sigma\frac{d}{d\sigma}\left(\frac{r(\sigma)}{f'(s(\sigma))}\right)\right)d\sigma ,$$

i.e.

(28)
$$\frac{dA}{dt} = \frac{1}{2}\, t^{-1}(t - t\tau)^{-1/2} \int_{t\tau}^{t} (Q_1(\sigma) + Q_2(\sigma))d\sigma$$

with

$$Q_1(\sigma) := \sigma\, \frac{r'(\sigma)}{f'(s(\sigma))}$$

$$Q_2(\sigma) := s'(\sigma)\Big(\frac{1}{2} - \sigma\, \frac{r(\sigma)f''(s(\sigma))}{f'(s(\sigma))^2}\Big) .$$

We estimate each term separately. By Lemma 2 and (9), we have

(29)
$$\int_{t\tau}^{t} |Q_1(\sigma)|d\sigma < \int_{t\tau}^{t} \sigma\, \frac{\sigma^{\alpha-1}|r|_\alpha}{2\beta^2\kappa\sqrt{\sigma} - c\sigma^{1/2+\alpha}}\, d\sigma <$$

$$\Big(\frac{1}{2}\,\frac{1}{1/2 + \alpha}\, \beta^{-2}\kappa^{-1} + c(t)\Big)t^{1/2+\alpha}(1 - \tau^{1/2+\alpha})|r|_\alpha .$$

We write Q_2 in the form

$$Q_2(\sigma) = \frac{r(\sigma)}{f'(s(\sigma))^3}\,\Big(\frac{1}{2}\,(f'(g(\sqrt{R(\sigma)})))^2 - \sigma r(\sigma)f''(s(\sigma))\Big) .$$

Since by (9), (10) and Lemma 2,

$$\left.\begin{array}{l} |\frac{1}{2}\,(f'(g(\sqrt{R(\sigma)})))^2 - 2\beta^2R(\sigma)| \\[2ex] |\sigma r(\sigma)f''(s(\sigma)) - 2\beta^2\sigma r(\sigma)| \end{array}\right\} < c\sigma^{3/2} ,$$

we obtain, using also (25),

(30)
$$\int_{t\tau}^{t} |Q_2(\sigma)|d\sigma < \int_{t\tau}^{t} \frac{r(0)}{(2\beta^2\kappa\sqrt{\sigma})^3}\,(1 + c(\sigma))\Big(2\beta^2\,\frac{1}{1 + \alpha}\,\sigma^{1+\alpha}|r|_\alpha\Big)d\sigma <$$

$$\Big(\frac{1}{4}\,\frac{1}{(1 + \alpha)(1/2 + \alpha)}\, \beta^{-2}\kappa^{-1} + c(t)\Big)t^{1/2+\alpha}(1 - \tau^{1/2+\alpha})|r|_\alpha .$$

Combining (29) and (30) with (28), it follows that

(31)
$$|(K_2r)(t)| < \frac{1}{4}\,\frac{1}{1/2 + \alpha}\, \beta^{-2}\kappa^{-1}\Big(1 + \frac{1}{2}\,\frac{1}{1 + \alpha}\Big)(1 + c(t))t^{\alpha-1}|r|_\alpha$$

$$\times\, \frac{1}{\sqrt{\pi}} \int_0^1 \frac{1 - \tau^{1/2+\alpha}}{(1 - \tau)^{3/2}}\, A_0(\tau)\exp(-A_0(\tau)^2)(1 + c(t))r(0)d\tau .$$

Adding the estimates (27) and (31) proves (21).

Remark. We conjecture that, for smooth initial data f, the function $r(t^2)$ is smooth, i.e.

$$r(t) = \kappa^2\beta^2 + r_{1/2}\sqrt{t} + r_1 t + \cdots . \tag{32}$$

Assuming an expansion of the form (32), we can calculate the coefficients $r_{1/2}, r_1, \ldots$ from the integral equation (15). In particular $f'''(0) \neq 0$ implies that $r_{1/2} \neq 0$. This shows that (8) is, in general, not valid for $\alpha > 1/2$.

REFERENCES

1. P. Benilan, M. G. Crandall, and A. Pazy, M-accretive operators, to appear.
2. L. C. Evans, Application of nonlinear semigroup theory to certain partial differential equations, in: Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, 1978.
3. A. Fasano and M. Primicerio, General free boundary problems for the heat equation, I, J. of Math. Anal. Appl. 57 (1977), 694-723.
4. A. Fasano and M. Primicerio, General free boundary problems for the heat equation, II, J. of Math. Anal. Appl. 58 (1977), 202-231.
5. K. Höllig, Existence of infinitely many solutions for a forward backward heat equation, Trans. Amer. Math. Soc., to appear.
6. K. Höllig and J. A. Nohel, A diffusion equation with a nonmonotone constitutive function, MRC Technical Summary Report #2443, Proceedings NATO/LONDON Math. Soc. Conference on Systems of Partial Differential Equations (Oxford U., August 1981), to appear.
7. D. Kinderlehrer and L. Nirenberg, Regularity in free boundary problems, Annali della SNS, 4 (1977), 373-391.
8. D. Schaeffer, A new proof of the infinite differentiability of the free boundary in the Stefan problem, J. Diff. Eq. 20 (1976), 266-269.

NONLINEAR INVERSE HEAT TRANSFER CALCULATIONS IN GUN BARRELS*

Alfred S. Carasso
Center for Applied Mathematics
National Bureau of Standards
Washington, DC 20234

ABSTRACT

We consider the problem of determining the temperature history inside a gun barrel from embedded thermocouple measurements at some distance away from the inside wall. This inverse problem leads to an improperly posed initial value problem for a nonlinear system of partial differential equations, whenever the thermal properties are temperature dependent. We discuss a step-by-step marching algorithm for the numerical computation of such problems. The scheme is stabilized by appropriately filtering in the frequency domain at each step. We illustrate this technique with a numerical experiment on a nonlinear problem whose exact solution is known. The basic ideas are applicable to other unstable evolution equations.

I. Introduction

This report summarizes the results of an important computational experiment on a nonlinear inverse heat conduction problem whose exact solution is known. We consider the problem of determining the temperature history at the inside wall of a gun barrel, from embedded thermocouple measurements at various points in the annular metallic region between the inner and outer radii of the cannon. As the shell is fired, a continuous trace is recorded at each thermocouple, providing temperature as a function of time at the corresponding fixed spatial location.

* Research sponsored by the U.S. Army Research Office under MIPR No. ARO 63-82

The present study centers around a novel computational technique designed especially for coping with the nonlinear case of temperature dependent thermal properties. It is a sequel to [1] where the linear quarter plane problem with constant coefficients, was thoroughly analyzed. As was shown there, in that case, the inverse problem can be formulated either as a Volterra integral equation of the first kind, or equivalently, as an initial value problem for the one dimensional heat equation run sideways. Either formulation leads to an improperly posed problem in which the solution, when it exists, depends discontinuously on the data.

The inverse problem can be regularized in the L^2 norm by placing an a-priori bound M on the norm of the unknown temperature history, $f(t)$, at the inside wall $x = 0$; at the same time, the measured noisy temperature data $g_m(t)$, at the location $x = \ell > 0$, is regarded as differing by at most ε in the L^2 norm from unknown smooth exact data $g(t)$, for which a solution exists. It is assumed that ε and M are known and compatible. As shown in [1, equations (2.20), (2.21)] this leads to explicit formulae for the temperature and gradient histories at each fixed x, $0 < x < \ell$. Also, error estimates are obtained for the regularized solutions implying Hölder continuity with respect to the data, for each fixed positive x. These estimates degenerate at the wall, [1, Theorem 1].

The regularization procedure can be interpreted as solving the initial value problem for the sideways heat equation with appropriately modified initial data. An explicit finite difference scheme consistent with that problem is shown to be unconditionally convergent, when used with the filtered initial data, [1, Theorem 3]. This step-by-step marching scheme in the x-variable is the basis for our approach to the nonlinear case of a temperature-dependent diffusion coefficient. We regularize the calculation at each step by filtering in the frequency domain, using FFT algorithms; we then return to the physical variables for the calculation of the next step. The filtering function used at each step is that determined by the related constant coefficient problem. This algorithm is outlined in [1, Section 7].

In order to test the robustness of this procedure, an example was manufactured with a known exact solution. This is a fictitious mathematical problem, artificially created so as to have a solution which simulates conditions presumed to exist in a 155mm cannon. The relevant parameters were made available to us by Dr. A. K. Celmins, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Grounds, Maryland. The "exact" solution was constructed numerically by solving a well posed direct problem as explained below.

2. The Direct Problem

Consider the initial boundary value problem

$$(2.1)\qquad \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}[a(u)\frac{\partial u}{\partial x}],\ 0 < x < \ell,\ t > 0,$$

$$(2.2)\qquad u(0,t) = f(t), \qquad u(1,t) = h(t),\ t > 0,$$

$$(2.3)\qquad u(x,0) = 300^\circ\ K$$

where t is the time measured in milliseconds, x measured in millimeters represents distance away from the inside wall, and u(x,t) is the temperature in degrees Kelvin. The heat conduction equation (2.1) is a simplification of the actual physical situation, in that first order terms arising from cylindrical symmetry have been neglected, as well as the variation of specific heat with temperature. Moreover, for gun steel at temperatures between 300° K and 1000° K, the conduction coefficient a(u) in (2.1) is well approximated by a linear function of u,

$$(2.4) \qquad a(u) = \{1.299 - 1.144 \times 10^{-3} (u - 255)\} \times 10^{-2} \text{ mm}^2/\text{millisec}$$

We remark that the methodology to be discussed can easily accommodate the more exact differential equation, as well as more complicated dependencies of a(u) on u . We shall refer to the quantity

$$(2.5) \qquad w(x,t) = -a(u) \frac{\partial u}{\partial x}$$

as the <u>temperature gradient</u>, by an abuse of te minology. It is measured in <u>mm° K/milliseconds</u>. In all Figures shown below dealing with plots of w(x,t) as a function of t for some fixed x , the vertical axis bears the legend "temperature gradient."

The functions f(t) and h(t) in (2.2) represent, respectively, the temperature histories at the inside wall and at 1mm away from the wall. These mathematical functions are plotted in Figure 1; they are constructed so as to approximate observed temperature histories in gun barrels, [4].

The direct problem given by (2.1), (2.2), and (2.3) was solved numerically, using an adaptive partial differential equation software package, MOL1D, [3]. The numerical integration was carried out to a distance in time equal to 100 milliseconds. The temperature $u(x,t)$ and gradient $w(x,t)$ were evaluated at various fixed values of x, as functions of time, and stored for subsequent comparisons. Figures 2 and 3 show the histories of u and w at $x = .25$mm. As is evident from Figure 3, the numerical calculation of w is not free from noise. Nevertheless, we use the term "exact solution" for any history obtained by the above numerical computation of the direct problem. All histories are records consisting of 400 equispaced samples on the time interval [0,100] milliseconds.

3. The Inverse Problem

The physical region of interest here is the x interval between 0 and .25mm. The histories in Figures 2 and 3 simulate what might have been recorded by a thermocouple at .25mm away from the inside wall as the shell is fired. The object is to use such data to reconstruct the temperature and gradient histories, arbitrarily close to the inside wall. In actuality, two thermocouple readings are necessary at $x = x_0$ and $x = x_1$, with $x_0 < .25 < x_1$; a well posed direct calculation, as in Section 2 above, then yields u and w at $x = .25$mm. As noted in the references given in [1], this type of inverse problem occurs in a variety of heat transfer contexts. The purpose of our computational experiment is as follows:

a) To demonstrate the feasibility of the inverse calculation in a realistic situation in which rapidly varying solutions and nonlinearity play a role. As may be seen from Figure 1, the postulated temperature at the wall rises from 300° K to almost 1000° K in the first 10 milliseconds. In this temperature range, the conduction coefficient a(u) given by (2.4) undergoes a 280 percent change.

b) To demonstrate the robustness of our algorithm with noisy data and a fine grid.

c) The regularized marching procedure we shall use is a powerful general method, applicable to other ill-posed evolutionary partial differential equations, linear and nonlinear. As used here, it is an adaptation to the nonlinear case of an algorithm which is rigorously justified in the constant coefficient case. While the heuristic "local mode analysis" underlying our regularization is likely to be valid in many other cases of ill-posed initial value problems, there is a need for well-documented realistic inverse calculations.

Let $z = \ell - x$ and let a_0, a_1 be positive constants such that

$$(3.1) \qquad 0 < a_0 \leqslant a(u) \leqslant a_1 .$$

Let $b(u) = \frac{da}{du}$, and let $v(z,t) = u(x,t)$. Using (2.5), we may write (2.1) as an equivalent first order system

$$(3.2) \qquad v_z = \frac{w}{a(v)} , \qquad w_z = v_t , \qquad 0 < z < \ell, \ t > 0,$$

with the subscript notation used for partial derivatives.

to be integrated in the direction of increasing z from $z = 0$ to $z = \ell$; we use the initial values given in Figures 2 and 3 and the following boundary conditions at $t = 0$,

$$(3.3) \qquad v(z,0) = 300° \text{ K}, \quad w(z,0) = 0.$$

Let Δz be the increment in the z-variable and let $\ell = (N + 1)\Delta z$. Let $v^n(t)$, $w^n(t)$ denote, respectively, $v(n\Delta z,t)$, $w(n\Delta z,t)$, for $0 < n \leqslant N + 1$. The following finite difference approximation is second order accurate and explicit,

$$(3.4) \qquad v^{n+1}(t) = v^n(t) + \frac{\Delta z\, w^n(t)}{a(v^n(t))} + \frac{\Delta z^2\, v_t^n(t)}{2a(v^n(t))} - \frac{\Delta z^2\, b(v^n(t))[w^n(t)]^2}{2a^3(v^n(t))} .$$

$$(3.5) \qquad w^{n+1}(t) = w^n(t) + \Delta z v_t^n(t) + \frac{\Delta z^2 w_t^n(t)}{2a(v^n(t))} - \frac{\Delta z^2\, b(v^n(t))\, w^n(t)\, v_t^n(t)}{2a^2(v^n(t))}$$

An effective way of implementing this scheme is to use cubic spline interpolation at the 400 equally spaced mesh points on the time interval $[0,T]$. Differentiating the spline function produces $O(\Delta t^3)$ accurate derivatives $v_t^n(t)$, $w_t^n(t)$ at these same mesh points, and hence $v^{n+1}(t)$, $w^{n+1}(t)$ from (3.4), (3.5). The next step is to stabilize this process by filtering each of these functions in the frequency domain. This is accomplished by dividing the kth Fourier coefficient by the precomputed weight λ_k, where

$$(3.6) \qquad \lambda_k = (1 + \omega^2 \exp [\ell \sqrt{\frac{2|k|\pi}{a_0 T}}])^{\Delta z/\ell}$$

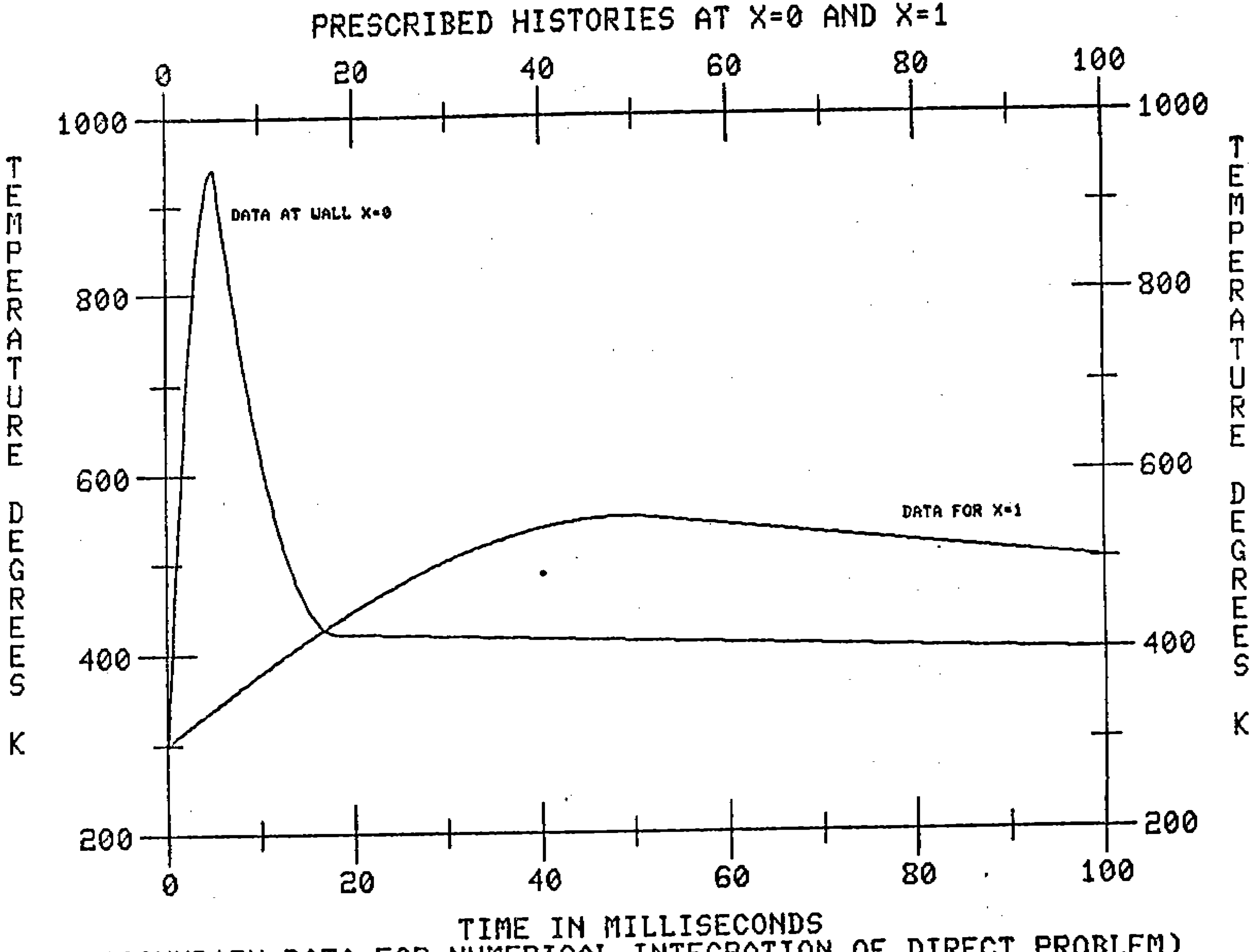
there $\omega = (\frac{\varepsilon}{M})$ is the L^2 noise to signal ratio. See [1].

With ℓ = .25mm, the x-interval $[0,\ell]$ was divided into 450 equally spaced mesh points, and the above procedure was implemented with ω = .001. Figures 4 through 11 summarize the comparison between exact and computed solutions at the interior location x = .056mm. An idea of the relative errors in the calculation is easily gained from Figures 7 and 11. Although the "logarithmic convexity" estimates in Theorem 1 of [1] degenerate at the wall, the computation was pursued for 450 cycles and approximations to the temperature and gradient histories at the wall were obtained. These are shown in Figures 13 and 17. The "exact" temperature and gradient histories at the wall are shown in Figures 12 and 16. Clearly, slight inaccuracies in the well-posed direct calculation of u(x,t) near x = 0, lead to a rather noisy determination of the exact w(x,t) at x = 0; in particular, the pronounced spike near t = 40 milliseconds in Figure 16 is a numerical artifact which should be disregarded. Nonetheless, we have chosen to compare the computed gradient history in Figure 17 with the wall profile given in Figure 16. As is evident from Figures 15 and 19, the wall estimates obtained by solving the inverse problem are quite reliable. This is especially true during the first twenty or so milliseconds where peak values are achieved.

4. Conclusions

A regularized marching algorithm has been shown to be effective in solving nonlinear inverse heat transfer problems in gun barrels. In [2], a similar technique was used successfully on linear backwards parabolic equations with highly variable coefficients. More recently, success has also been achieved on other unstable examples involving Burgers' equation with the time direction reversed.

Future work should be directed towards problems in two or more space dimensions in general domains, in the context of heat transfer and fluid mechanics.



(BOUNDARY DATA FOR NUMERICAL INTEGRATION OF DIRECT PROBLEM)

FIGURE 1

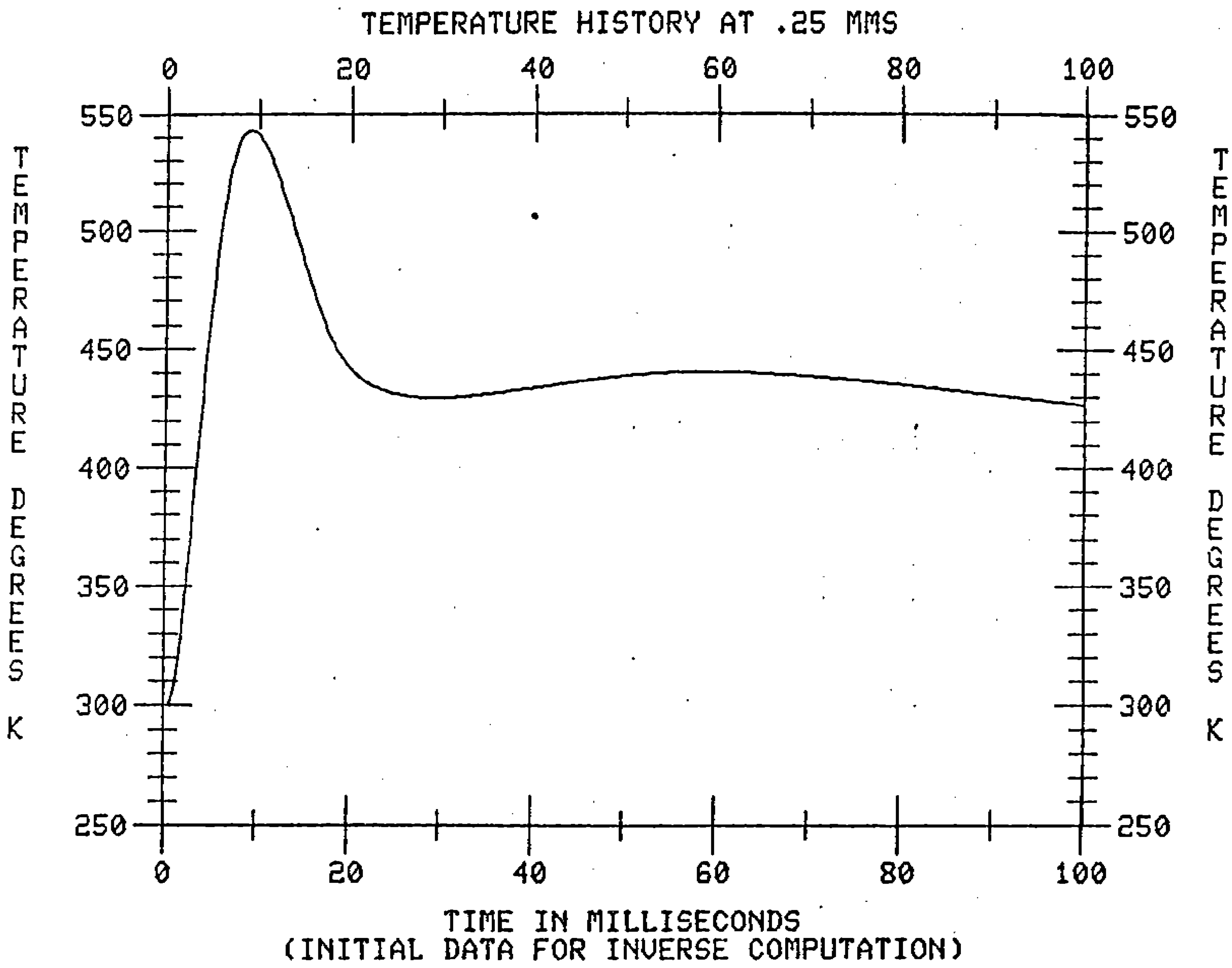

TEMPERATURE HISTORY AT .25 MMS
0
20
40
60
80
100
550
500
450
400
350
300
250
TEMPERATURE DEGREES K
TEMPERATURE DEGREES K
TIME IN MILLISECONDS
(INITIAL DATA FOR INVERSE COMPUTATION)

FIGURE 2

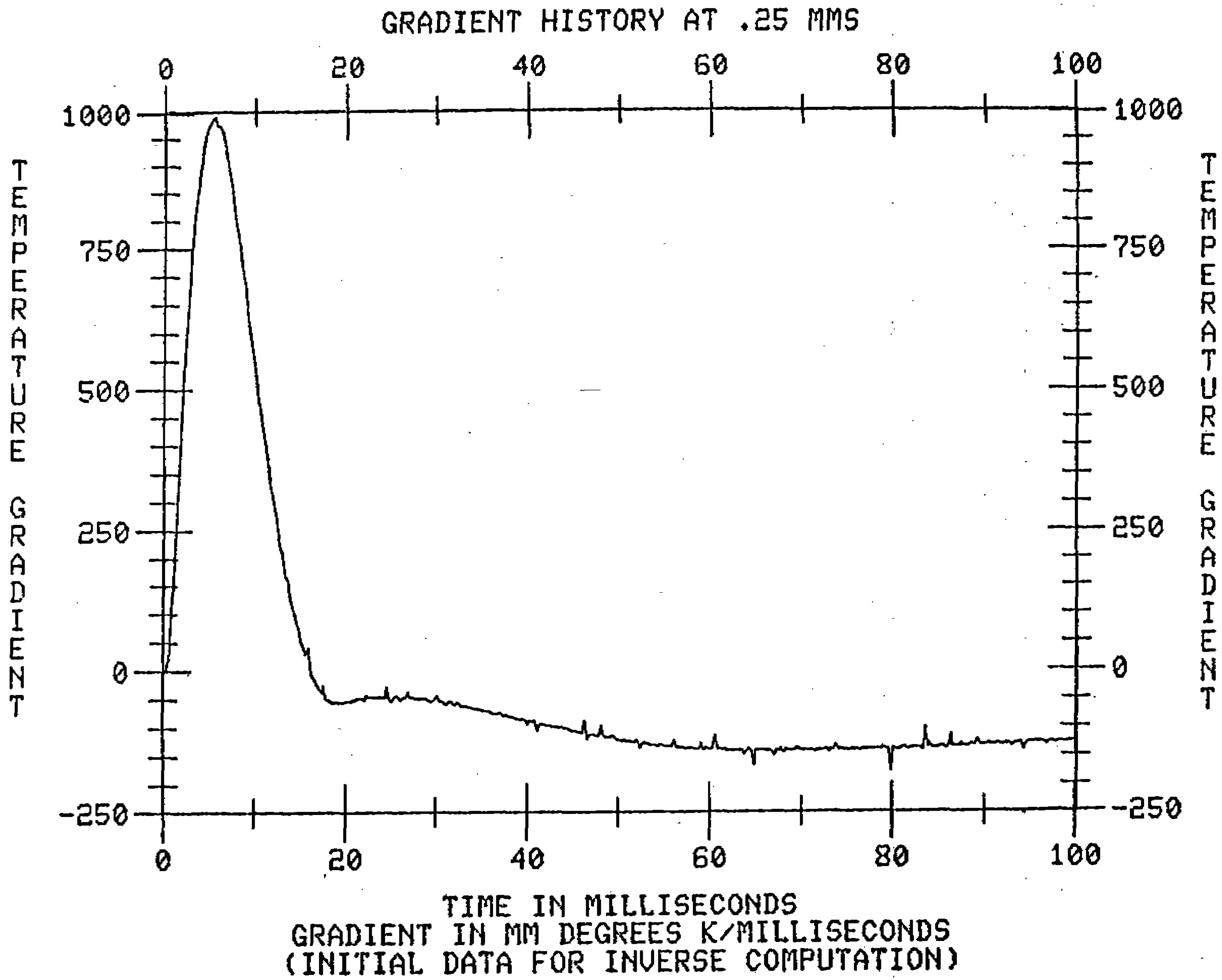

GRADIENT HISTORY AT .25 MMS
0
20
40
60
80
100
1000
750
500
250
0
-250
TEMPERATURE GRADIENT
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
(INITIAL DATA FOR INVERSE COMPUTATION)

FIGURE 3

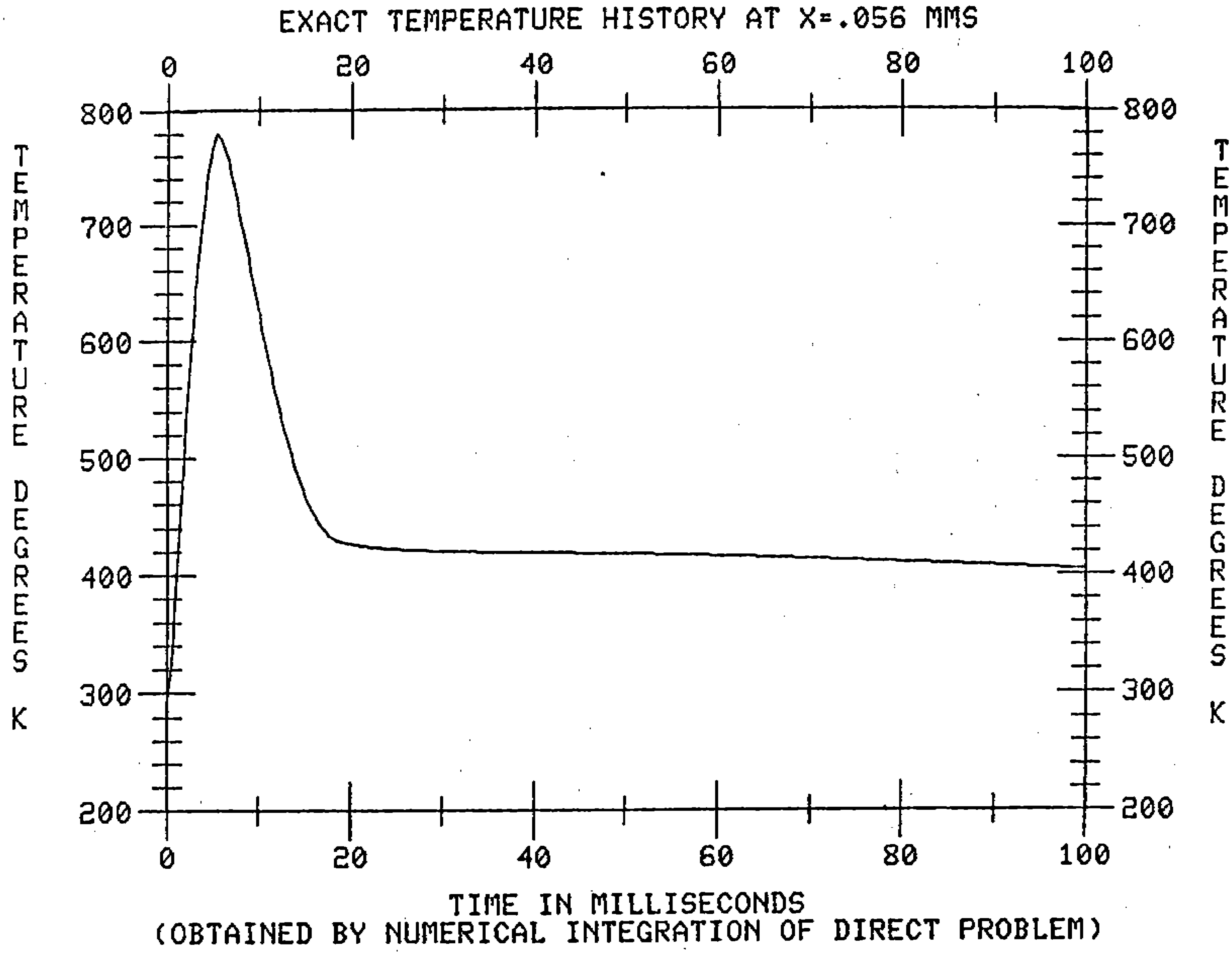

EXACT TEMPERATURE HISTORY AT X=.056 MMS
0
20
40
60
80
100
800
700
600
500
400
300
200
TEMPERATURE DEGREES K
TIME IN MILLISECONDS
(OBTAINED BY NUMERICAL INTEGRATION OF DIRECT PROBLEM)

FIGURE 4

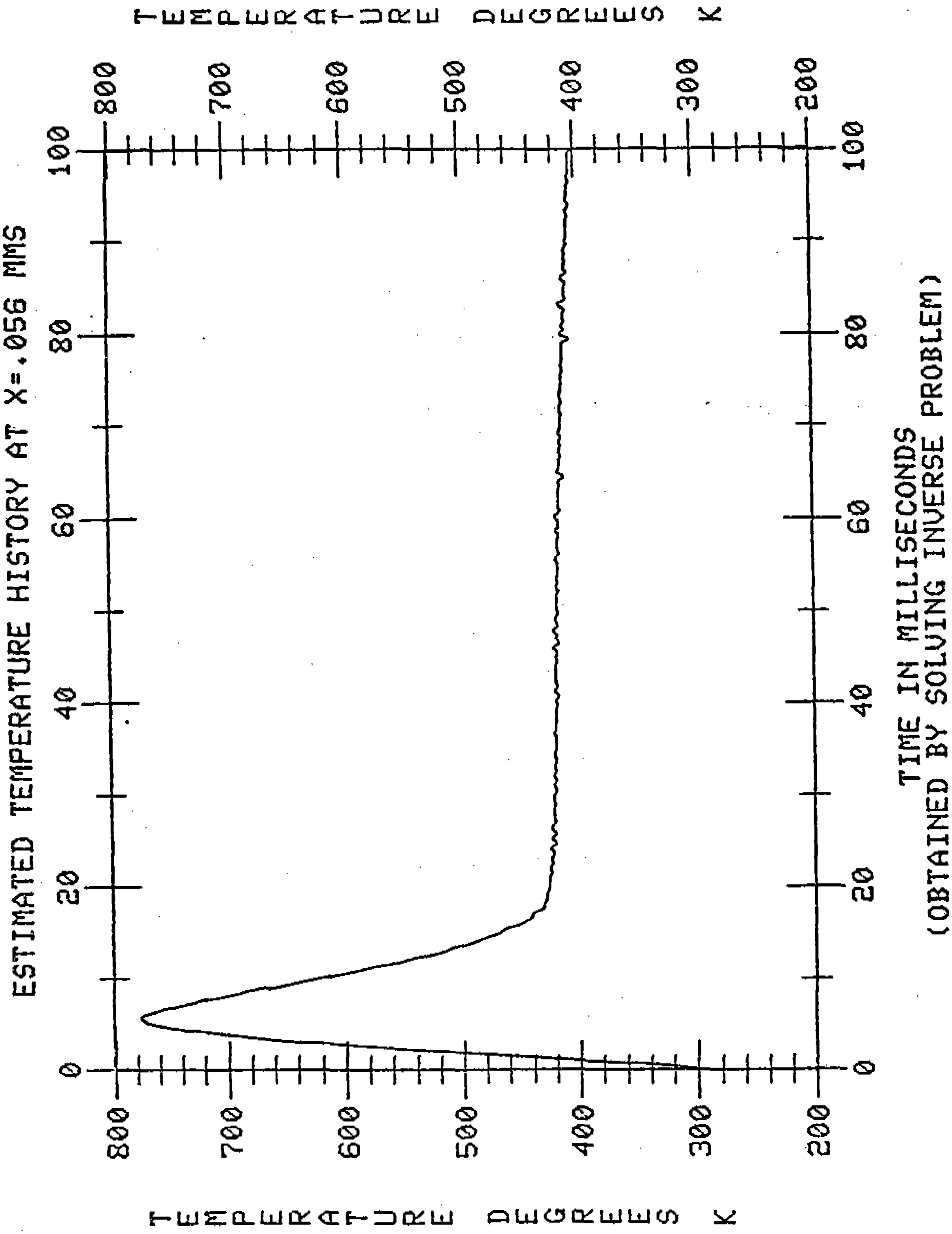

ESTIMATED TEMPERATURE HISTORY AT X=.056 MMS
TIME IN MILLISECONDS
(OBTAINED BY SOLVING INVERSE PROBLEM)
TEMPERATURE DEGREES K
TEMPERATURE DEGREES K
0
20
40
60
80
100
800
700
600
500
400
300
200

FIGURE 5

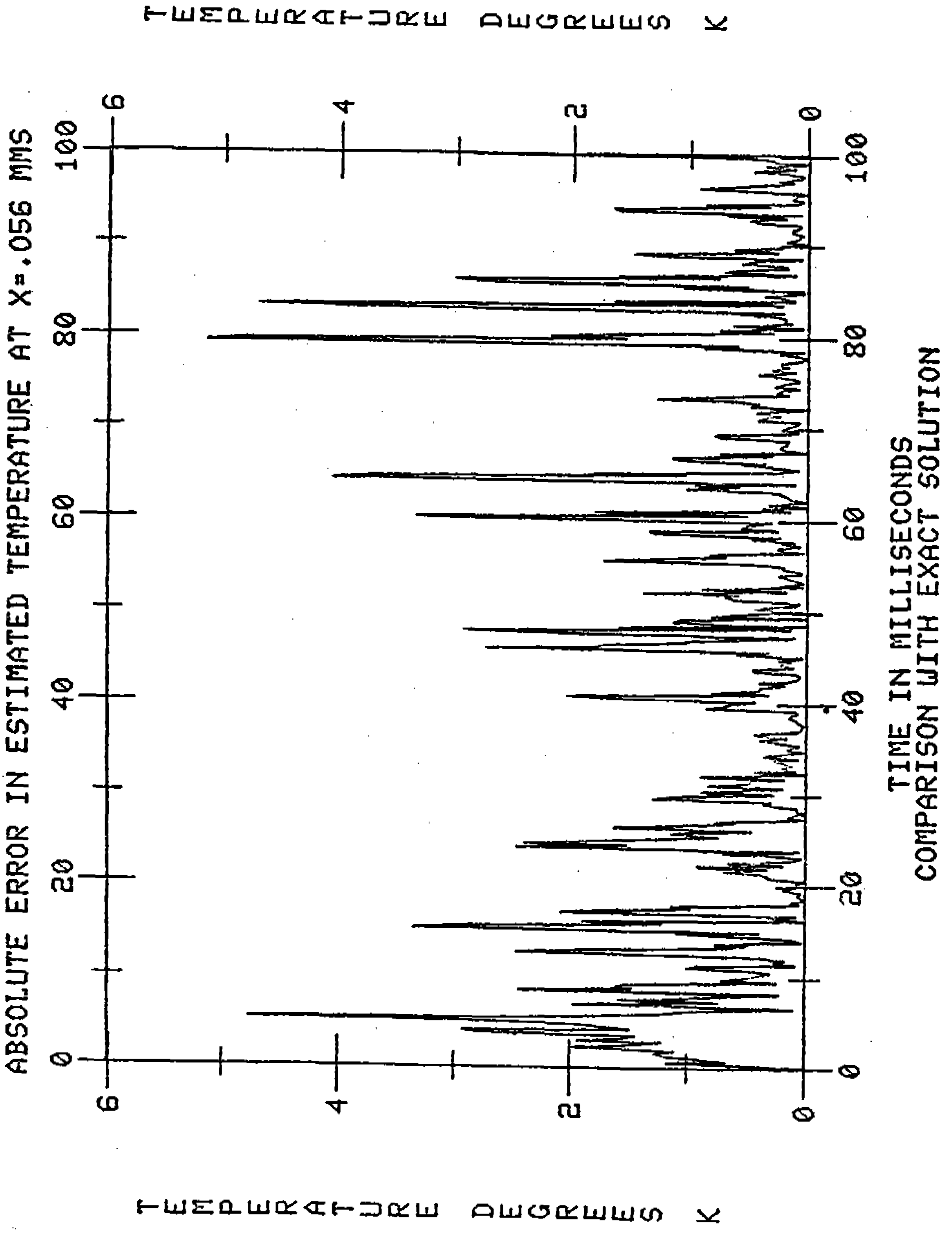

ABSOLUTE ERROR IN ESTIMATED TEMPERATURE AT X=.056 MMS
TIME IN MILLISECONDS
COMPARISON WITH EXACT SOLUTION
TEMPERATURE DEGREES K
TEMPERATURE DEGREES K
0
20
40
60
80
100
0
2
4
6

FIGURE 6

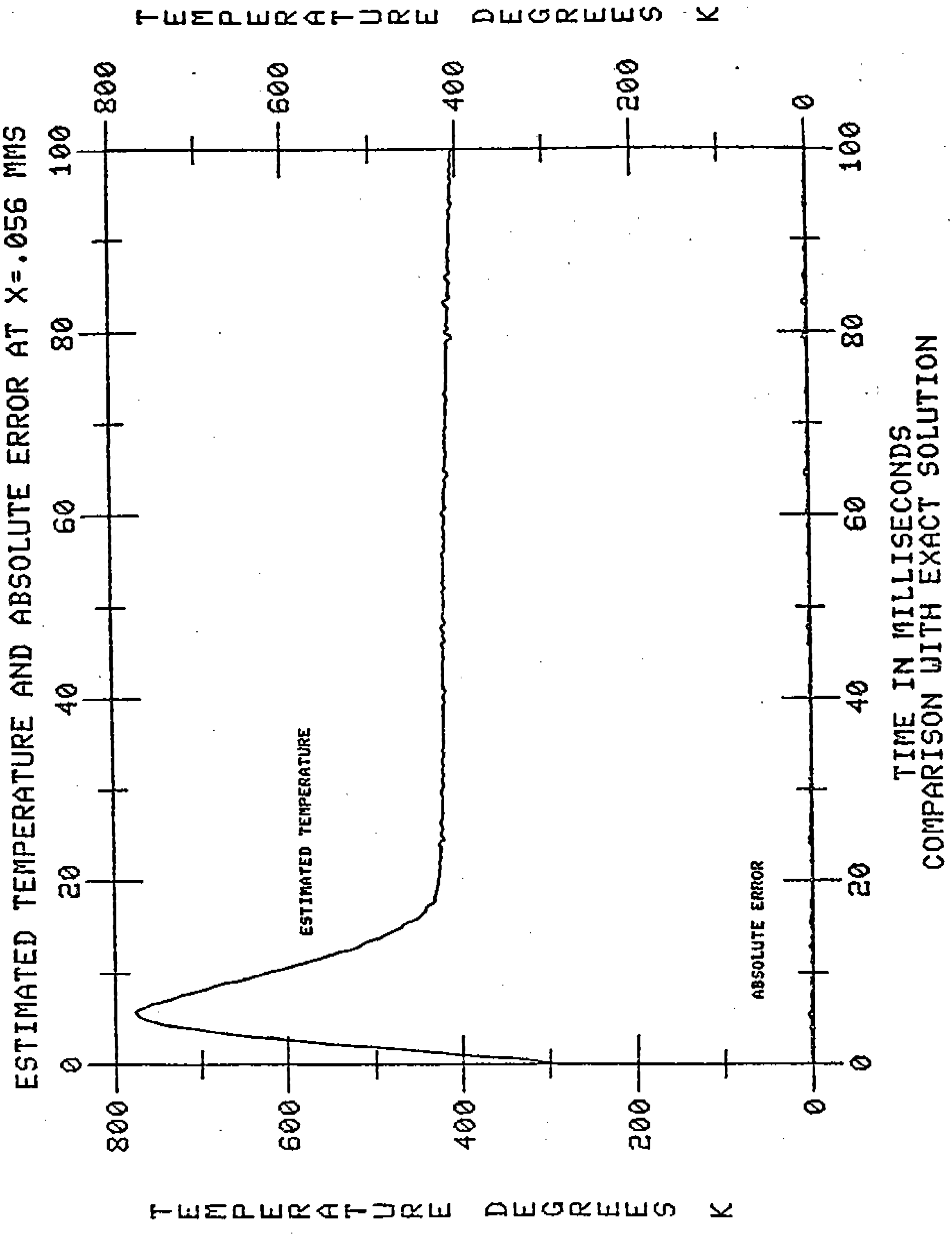

ESTIMATED TEMPERATURE AND ABSOLUTE ERROR AT X=.056 MMS
0
20
40
60
80
100
800
600
400
200
0
TEMPERATURE DEGREES K
TEMPERATURE DEGREES K
ESTIMATED TEMPERATURE
ABSOLUTE ERROR
TIME IN MILLISECONDS
COMPARISON WITH EXACT SOLUTION

FIGURE 7

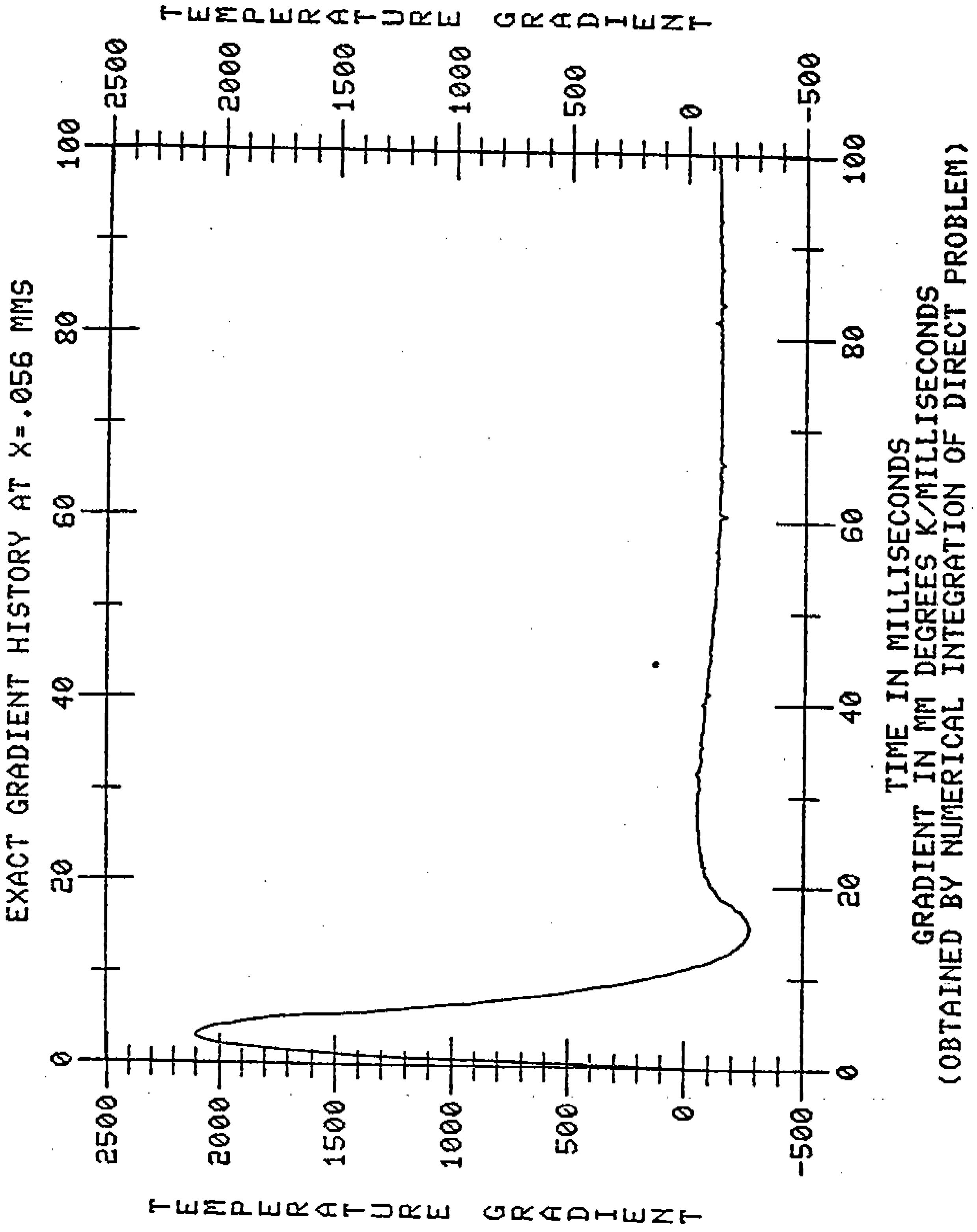

EXACT GRADIENT HISTORY AT X=.056 MMS
0
20
40
60
80
100
2500
2000
1500
1000
500
0
-500
TEMPERATURE GRADIENT
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
(OBTAINED BY NUMERICAL INTEGRATION OF DIRECT PROBLEM)

FIGURE 8

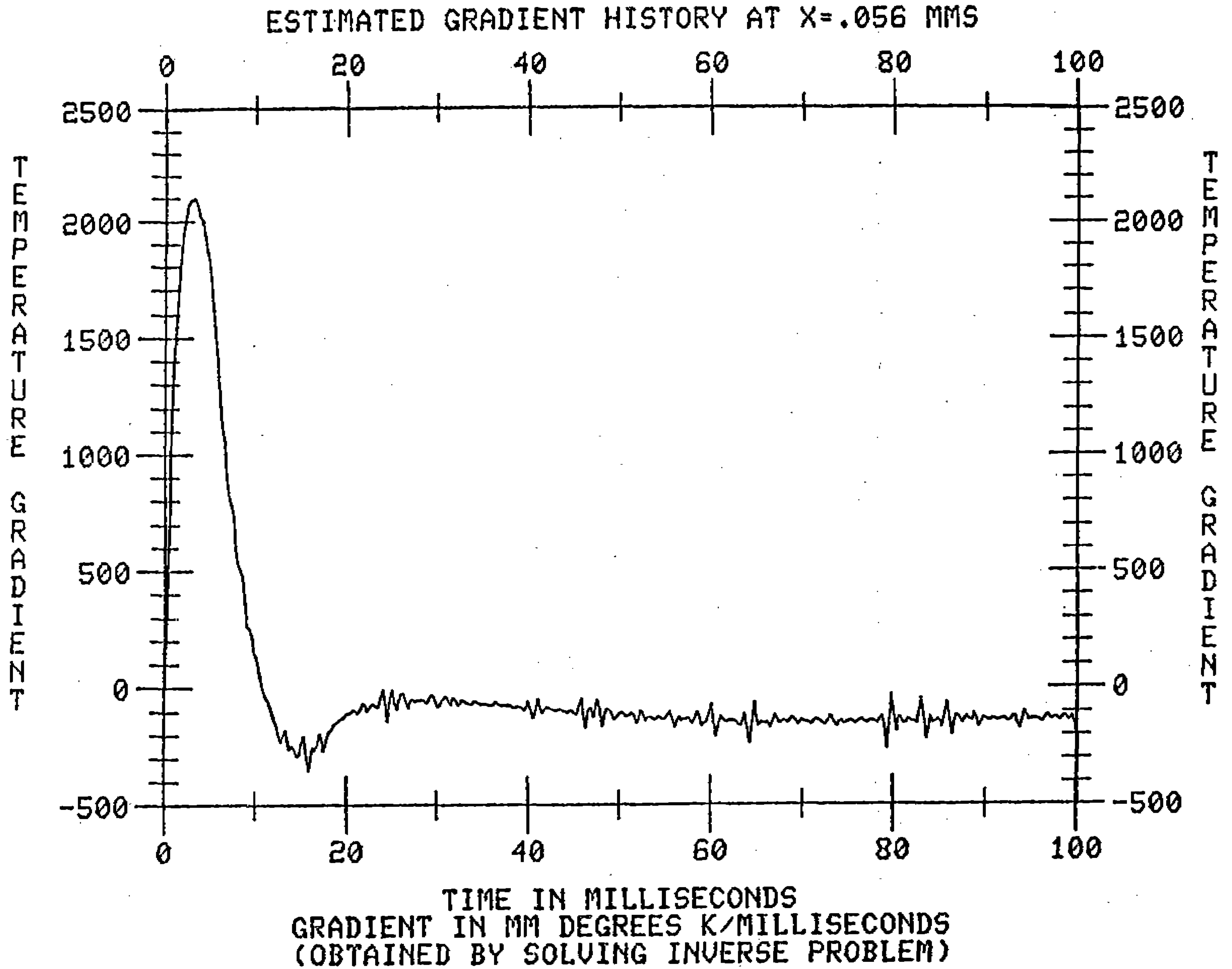

ESTIMATED GRADIENT HISTORY AT X=.056 MMS
0
20
40
60
80
100
2500
2000
1500
1000
500
0
-500
TEMPERATURE GRADIENT
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
(OBTAINED BY SOLVING INVERSE PROBLEM)

FIGURE 9

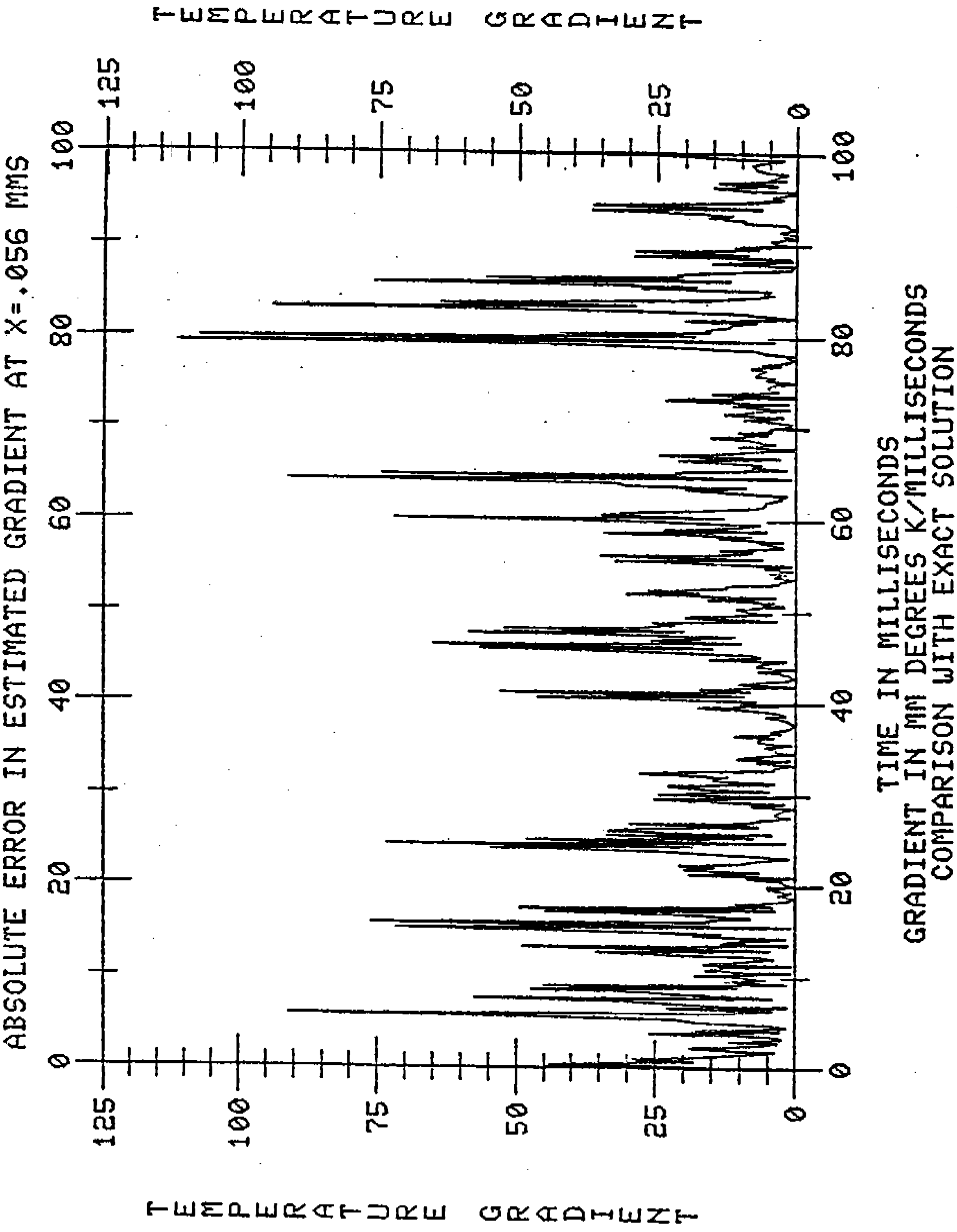

ABSOLUTE ERROR IN ESTIMATED GRADIENT AT X=.056 MMS
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
COMPARISON WITH EXACT SOLUTION
TEMPERATURE GRADIENT
TEMPERATURE GRADIENT
0
20
40
60
80
100
0
20
40
60
80
100
125
100
75
50
25
0
125
100
75
50
25
0

FIGURE 10

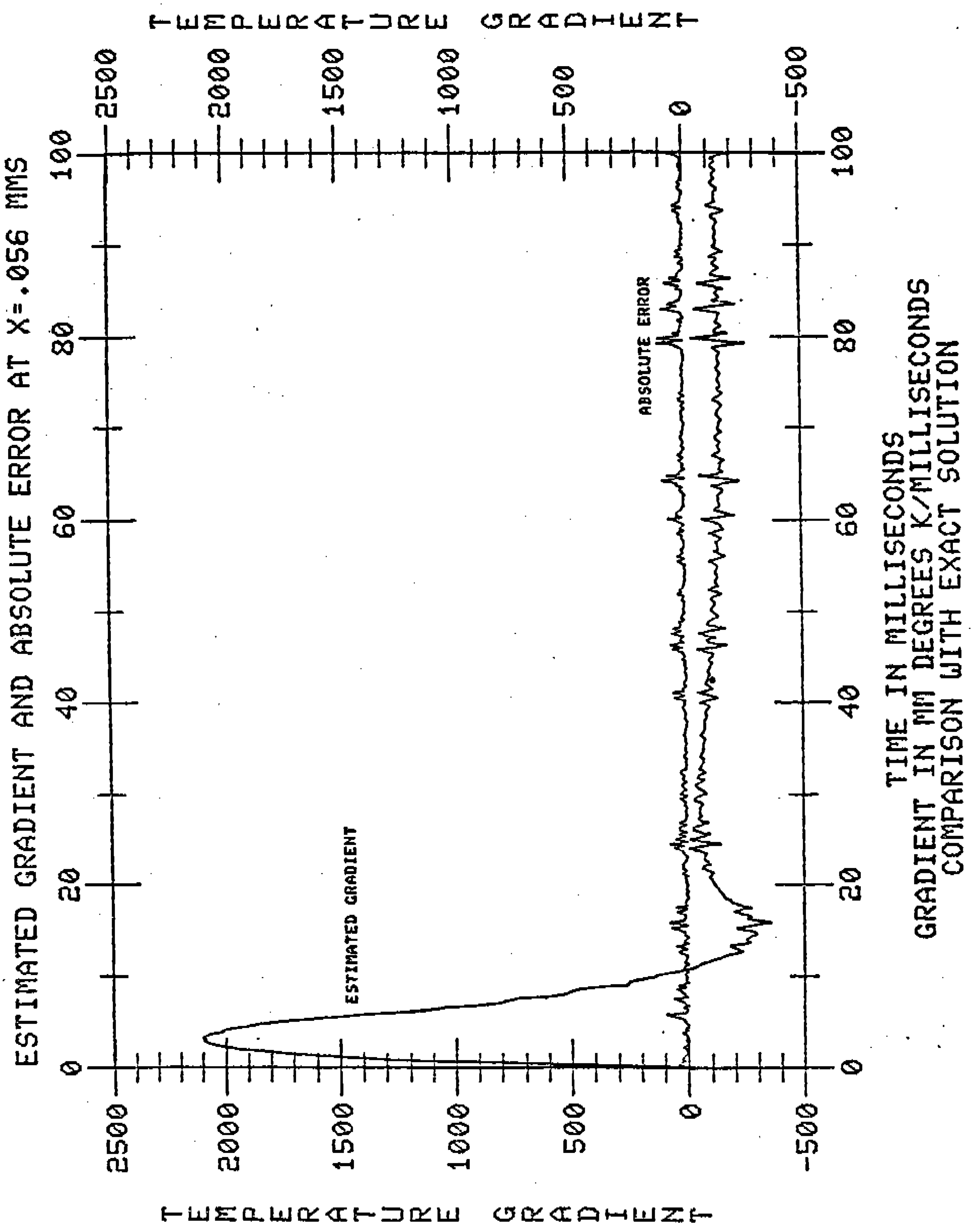

ESTIMATED GRADIENT AND ABSOLUTE ERROR AT X=.056 MMS
TEMPERATURE GRADIENT
2500
2000
1500
1000
500
0
-500
TEMPERATURE GRADIENT
0
20
40
60
80
100
ESTIMATED GRADIENT
ABSOLUTE ERROR
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
COMPARISON WITH EXACT SOLUTION

FIGURE 11

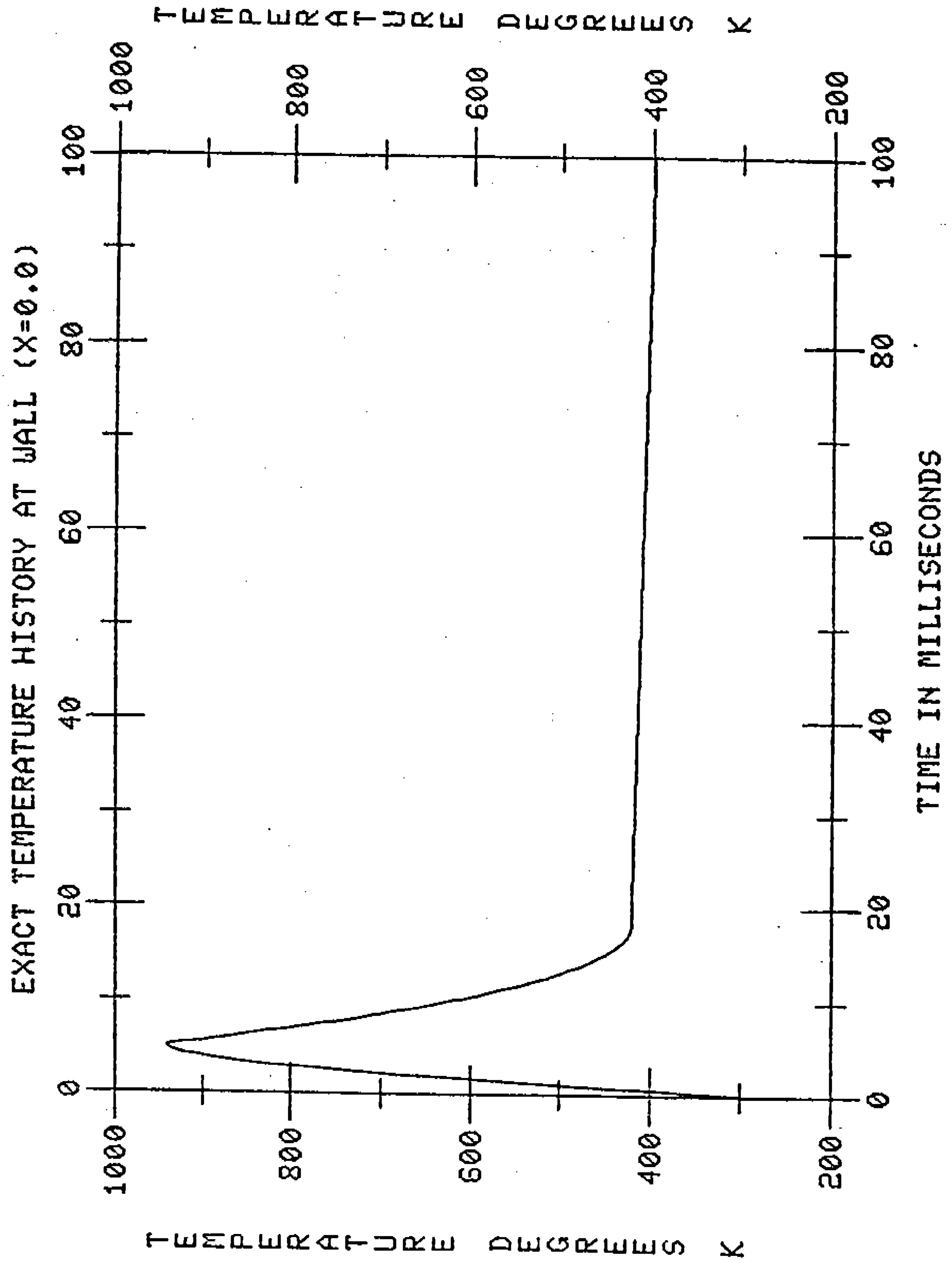

EXACT TEMPERATURE HISTORY AT WALL (X=0.0)
TEMPERATURE DEGREES K
1000
800
600
400
200
0
20
40
60
80
100
TIME IN MILLISECONDS
TEMPERATURE DEGREES K
1000
800
600
400
200

FIGURE 12

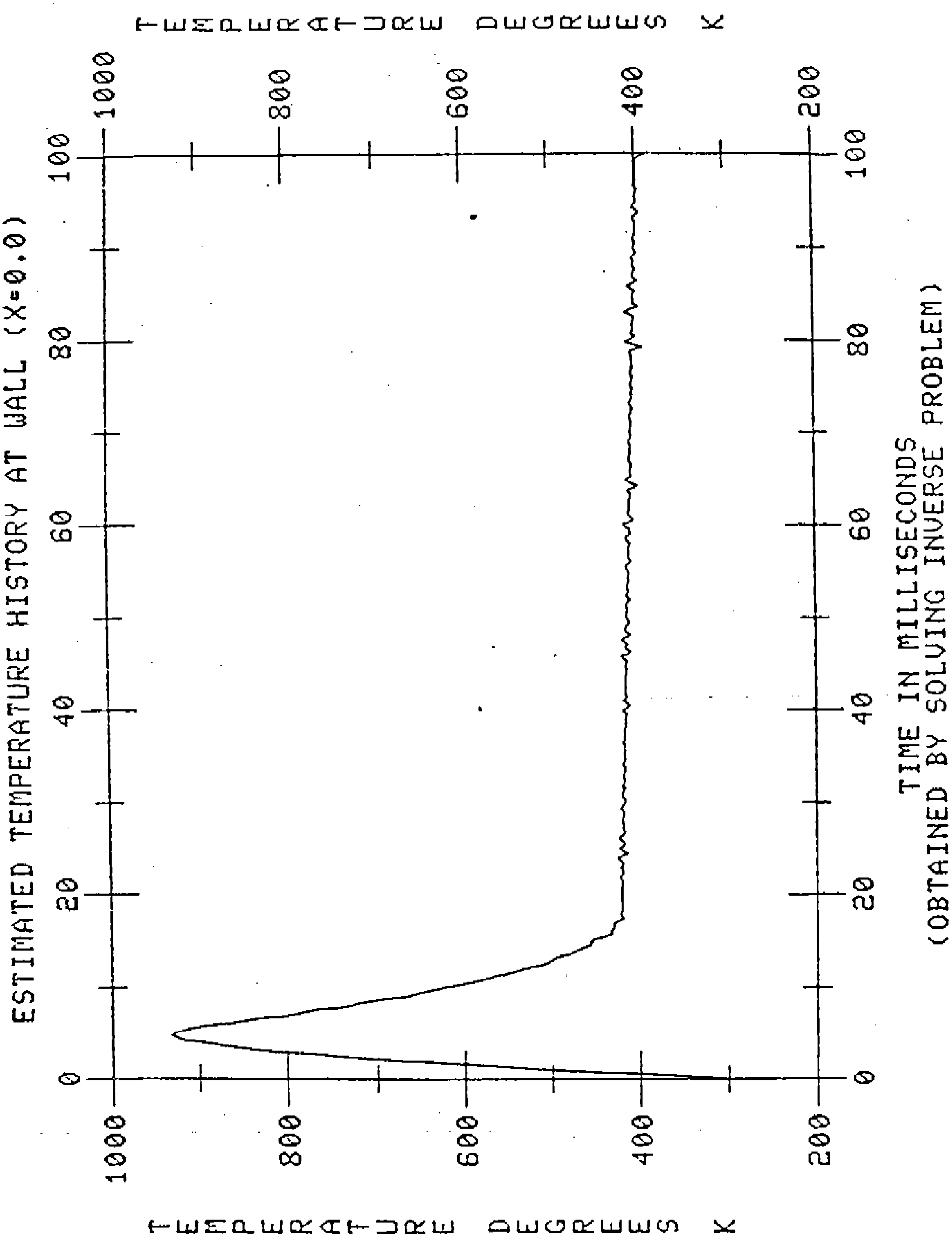

ESTIMATED TEMPERATURE HISTORY AT WALL (X=0.0)
TEMPERATURE DEGREES K
1000
800
600
400
200
0
20
40
60
80
100
TIME IN MILLISECONDS
(OBTAINED BY SOLVING INVERSE PROBLEM)
TEMPERATURE DEGREES K

FIGURE 13

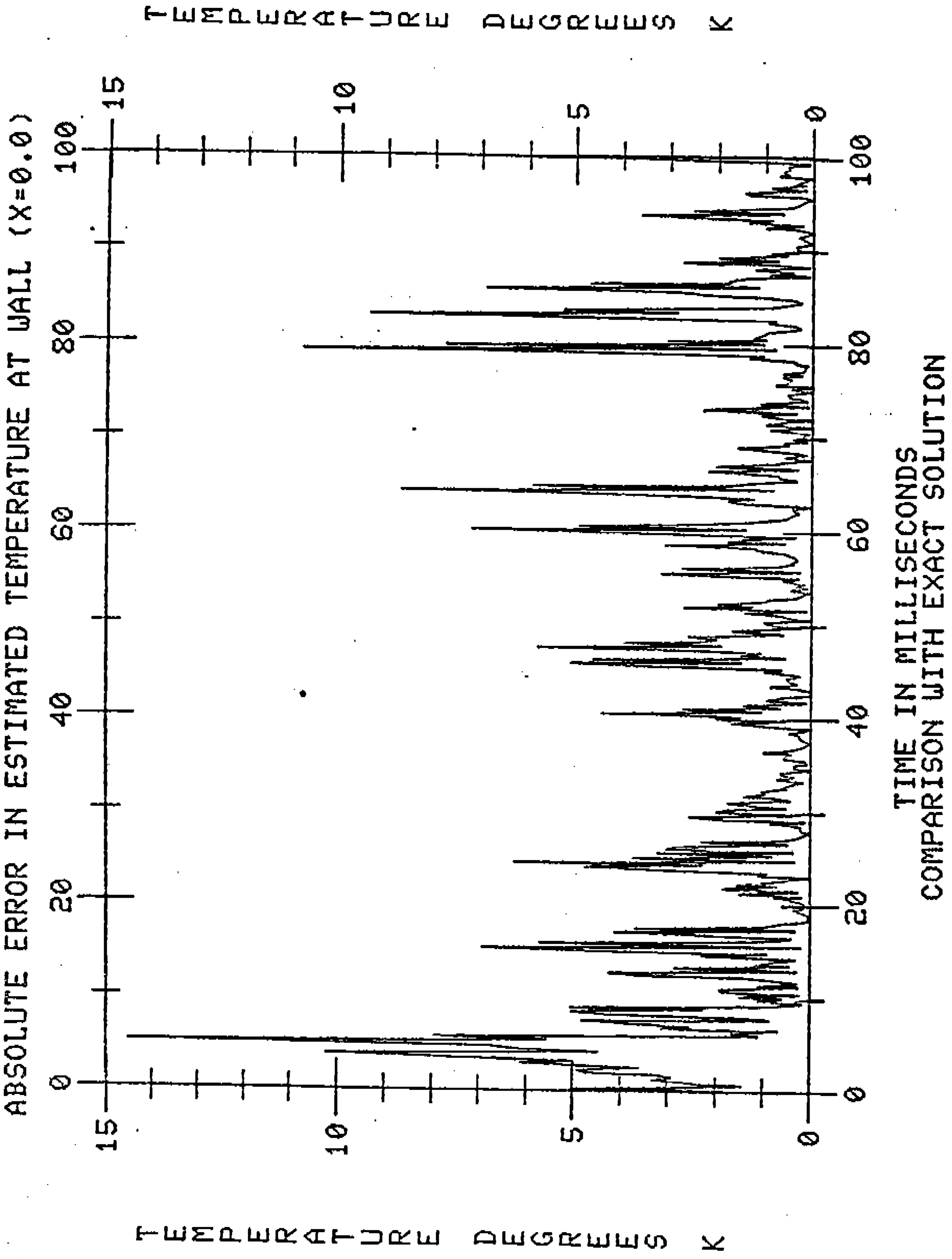

ABSOLUTE ERROR IN ESTIMATED TEMPERATURE AT WALL (X=0.0)
TEMPERATURE DEGREES K
15
10
5
0
0
20
40
60
80
100
TIME IN MILLISECONDS
COMPARISON WITH EXACT SOLUTION
TEMPERATURE DEGREES K

FIGURE 14

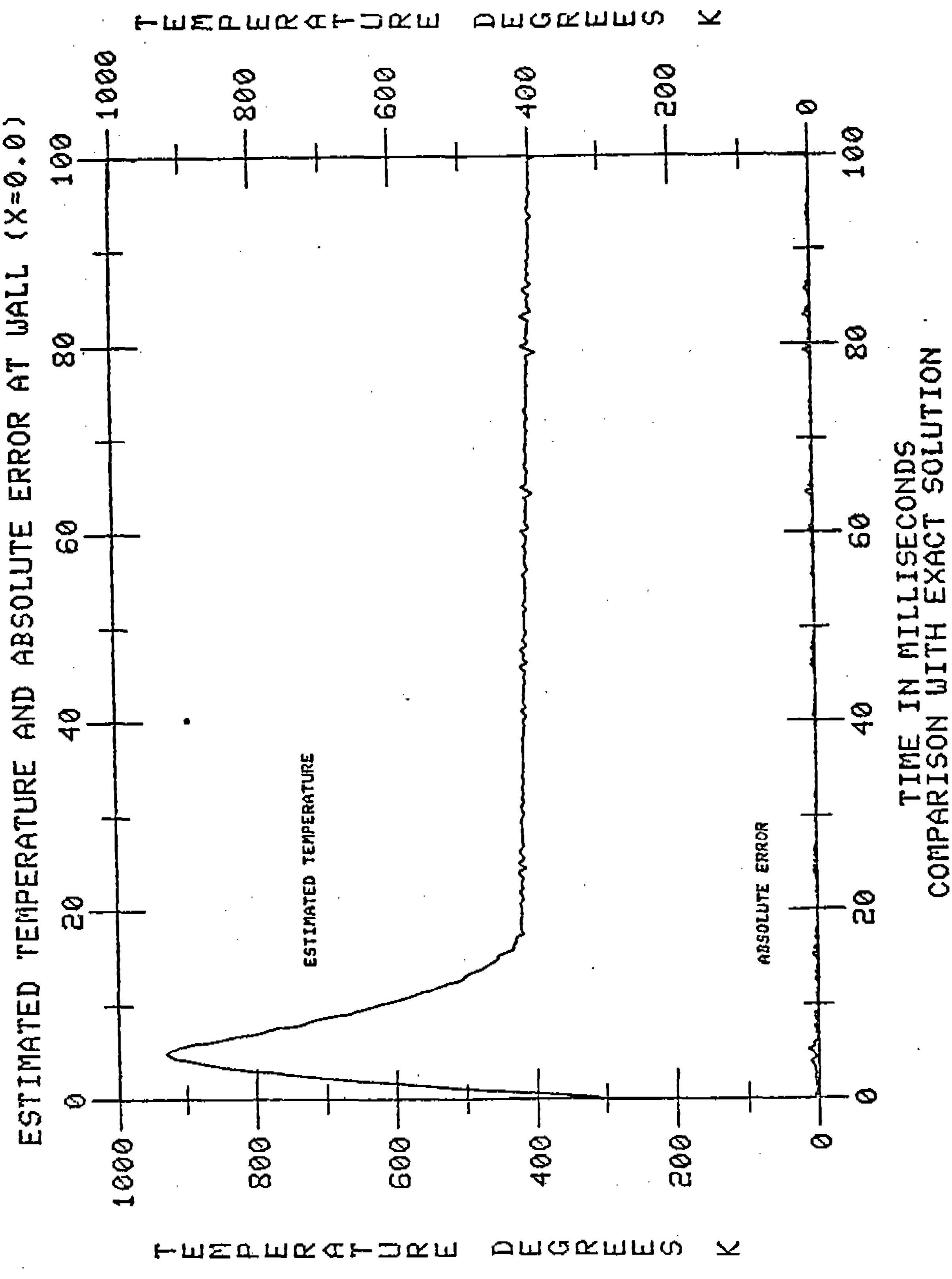

ESTIMATED TEMPERATURE AND ABSOLUTE ERROR AT WALL (X=0.0)
TEMPERATURE DEGREES K
TEMPERATURE DEGREES K
TIME IN MILLISECONDS
COMPARISON WITH EXACT SOLUTION
ESTIMATED TEMPERATURE
ABSOLUTE ERROR
0
20
40
60
80
100
0
200
400
600
800
1000

FIGURE 15

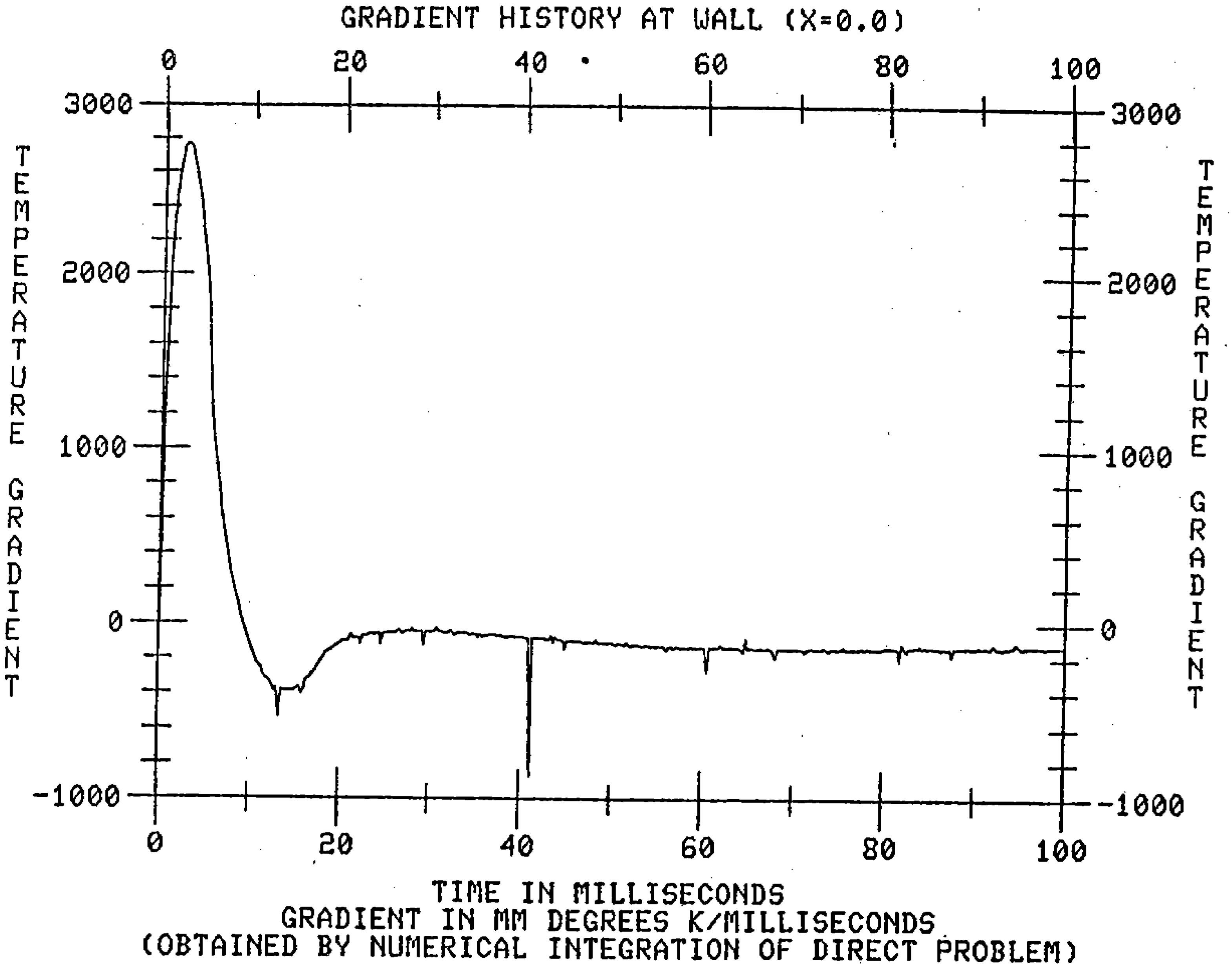

GRADIENT HISTORY AT WALL (X=0.0)
0
20
40
60
80
100
3000
2000
1000
0
-1000
TEMPERATURE GRADIENT
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
(OBTAINED BY NUMERICAL INTEGRATION OF DIRECT PROBLEM)

FIGURE 16

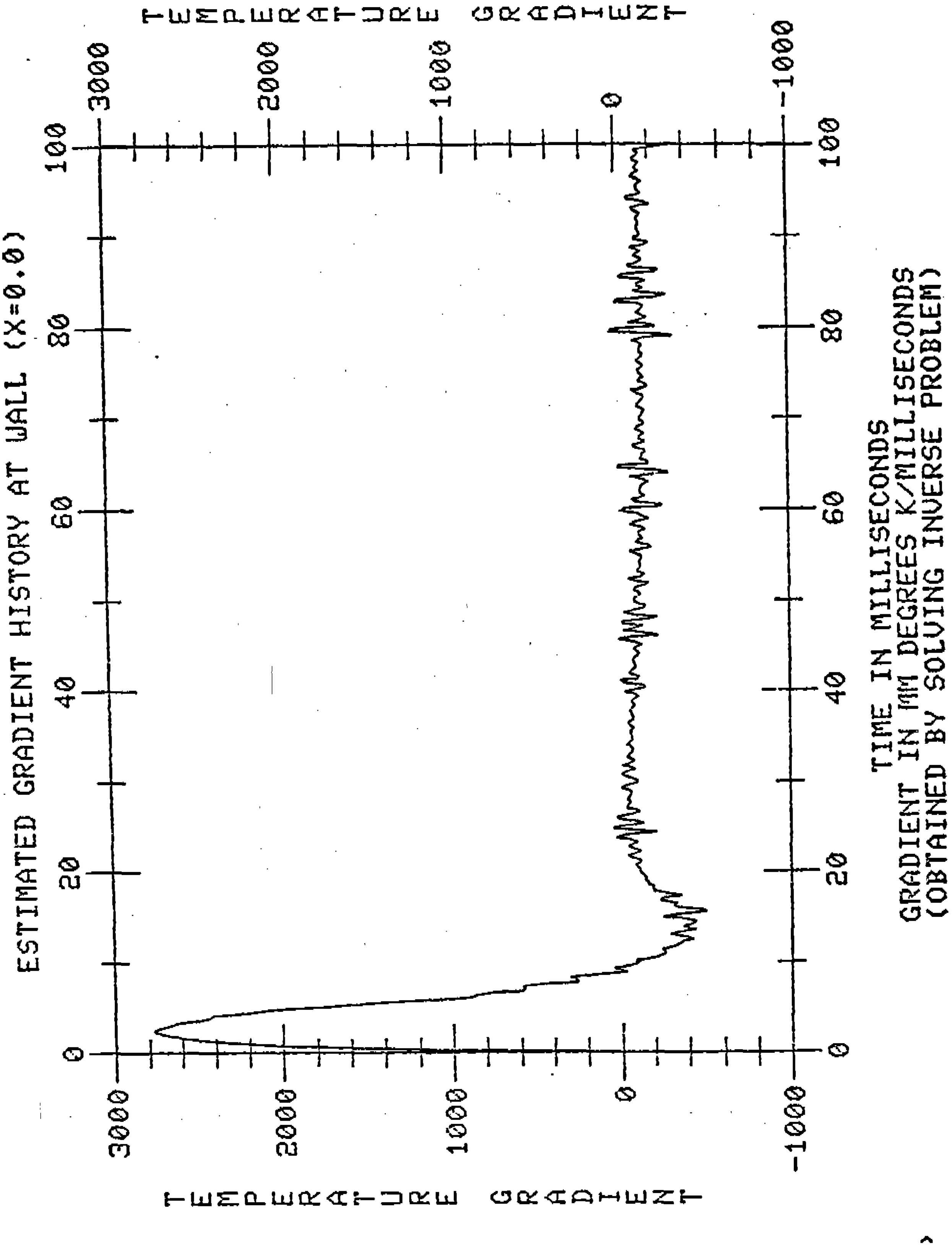

ESTIMATED GRADIENT HISTORY AT WALL (X=0.0)
0
20
40
60
80
100
3000
2000
1000
0
-1000
TEMPERATURE GRADIENT
3000
2000
1000
0
-1000
TEMPERATURE GRADIENT
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
(OBTAINED BY SOLVING INVERSE PROBLEM)

FIGURE 17

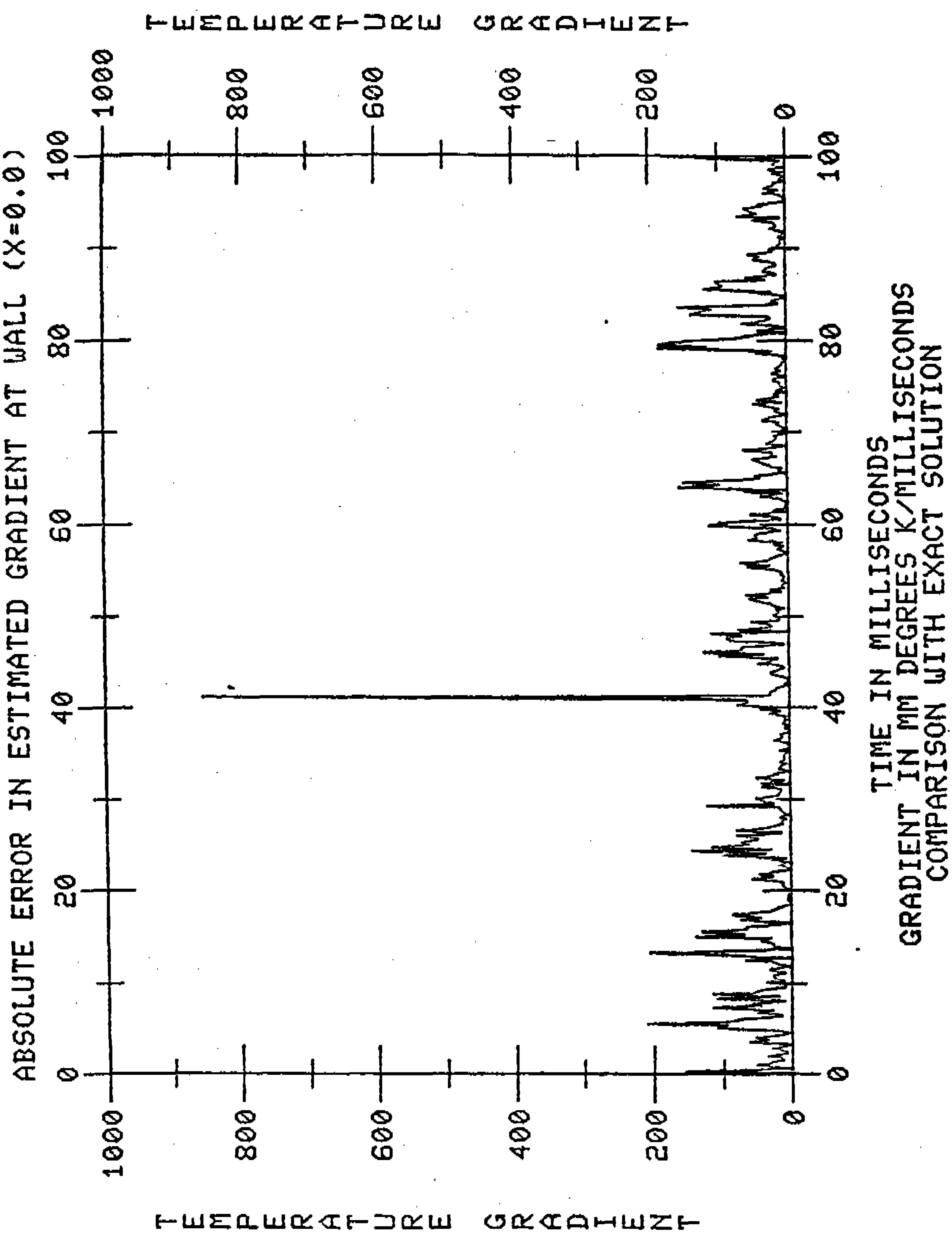

ABSOLUTE ERROR IN ESTIMATED GRADIENT AT WALL (X=0.0)
TEMPERATURE GRADIENT
1000
800
600
400
200
0
0
20
40
60
80
100
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
COMPARISON WITH EXACT SOLUTION

FIGURE 18

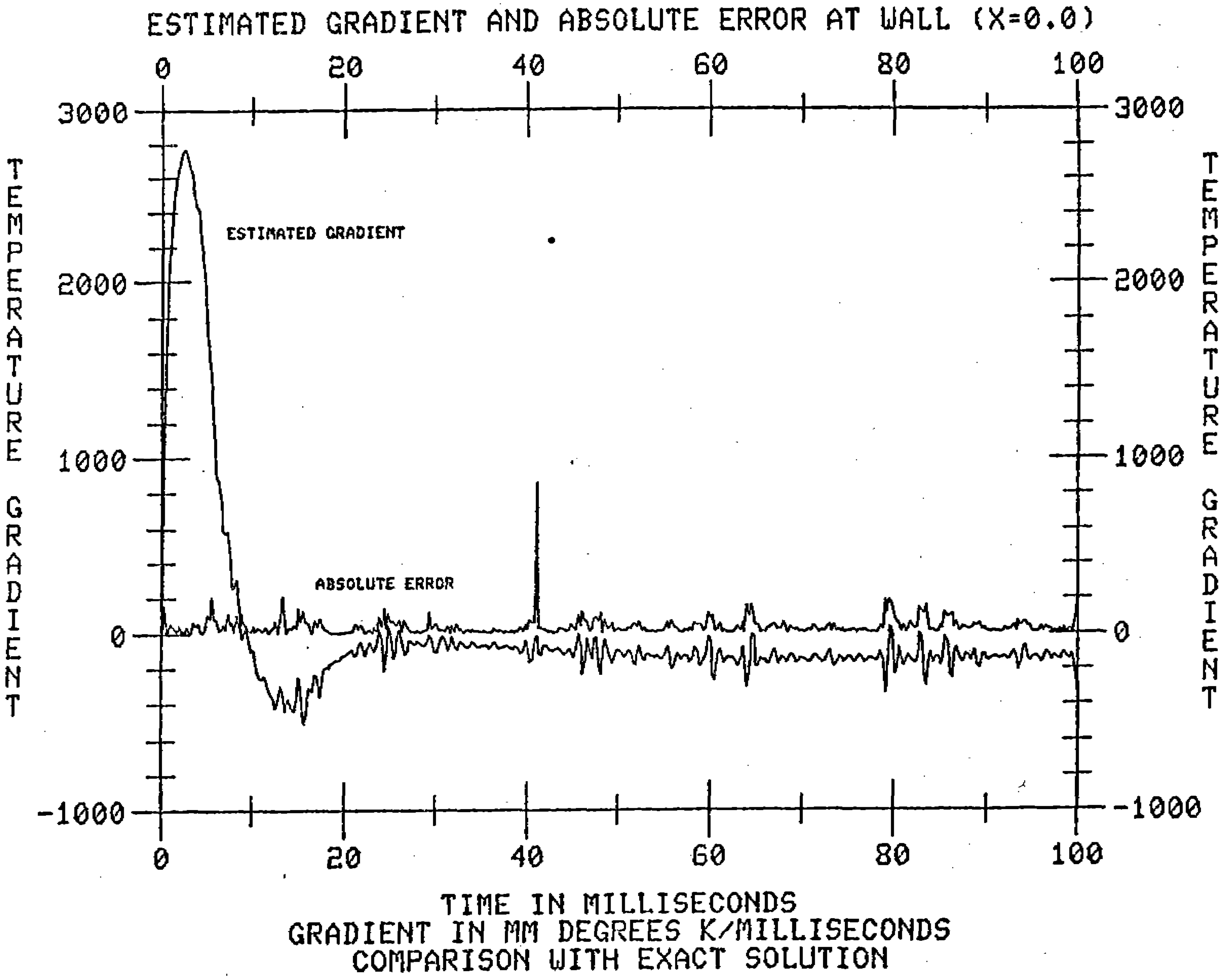

ESTIMATED GRADIENT AND ABSOLUTE ERROR AT WALL (X=0.0)
0
20
40
60
80
100
3000
2000
1000
0
-1000
TEMPERATURE GRADIENT
ESTIMATED GRADIENT
ABSOLUTE ERROR
TIME IN MILLISECONDS
GRADIENT IN MM DEGREES K/MILLISECONDS
COMPARISON WITH EXACT SOLUTION

FIGURE 19

REFERENCES

1. A. Carasso, "Determining Surface Temperatures from Interior Observations," SIAM J. Appl. Math., 42, (1982), pp. 558-574.

2. A. Carasso, "A Stable Marching Scheme for an Ill-Posed Initial Value Problem," Proceedings of the Oberwolfach Conference on Improperly Posed Problems and Their Numerical Treatment, Oberwolfach, Germany, (September 1982), K. H. Hoffmann, editor. International Series in Numerical Mathematics, Birkhauser-Verlag, Basel.

3. J. M. Hyman, "MOL1D: A General Purpose Subroutine Package for the Numerical Solution of Partial Differential Equations," LANL Report LA 7595M-UC32, (March 1979), Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

4. J. R. Ward and T. L. Brosseau, "Effect of Wear Reducing Additives on Heat Transfer into the 155mm M185 Cannon," BRL-M Report 2730, (February 1977), National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22161.

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

IMPLICATIONS OF ANALYTICAL INVESTIGATIONS ABOUT THE SEMICONDUCTOR EQUATIONS ON DEVICE MODELING PROGRAMS

Ch. Ringhofer* and S. Selberherr**

ABSTRACT

This paper gives guidelines for the development of computer programs for the numerical simulation of semiconductor devices. For this purpose the basic mathematical results on the corresponding elliptic boundary value problem are reviewed. In particular, existence, smoothness and structure of the solutions of the fundamental semiconductor equations are discussed. Various feasible approaches to the numerical solution of the semiconductor equations are described. Much emphasis is placed on constructive remarks to help authors of device simulation programs to make decisions on their code design problems. In particular, criteria for an optimal mesh generation strategy are given. The iterative solution of the systems of nonlinear and linear equations obtained by discretising the semiconductor equations is discussed. An example is given showing the power of these concepts combined with modern numerical methods in comparison to classical approaches.

AMS (MOS) Subject Classifications: 35J55, 35J60, 65N05, 65N10

Key Words:

*Mathematics Research Center, University of Wisconsin, Madison, WI 53705, USA.
**Abteilung Physikalische Elektronik, Institut fuer Allgemeine Elektrotechnik und Elektronik, TU Wien, Guszhausstrasze 27, A-1040, Wien, AUSTRIA.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2.

IMPLICATIONS OF ANALYTICAL INVESTIGATIONS ABOUT THE SEMICONDUCTOR EQUATIONS ON DEVICE MODELING PROGRAMS

Ch. Ringhofer* and S. Selberherr**

1. Introduction

The characteristic feature of early device modeling is the separation of the interior of the device into different regions, the treatment of which could be simplified by various assumptions like special doping profiles, complete depletion and quasineutrality. These separately treated regions were simply put together to produce the overall solution. If results in an analytically closed form are intended, any other approach is prohibitive. Fully numerical modeling based on partial differential equations /61/ which describe all different regions of semiconductor devices in one unified manner was first suggested by Gummel /29/ for the one dimensional bipolar transistor. This approach was further developed and applied to pn-junction theory by De Mari /13/, /14/ and to IMPATT diodes by Scharfetter and Gummel /50/.

A two dimensional numerical analysis of a semiconductor device was carried out first by Kennedy and O'Brien /35/ who investigated the junction field effect transistor. Since then two dimensional modeling has been applied to fairly all important semiconductor devices. There are so many papers of excellent repute that it would be unfair to cite only a few. Recently also the first results on three dimensional device modeling have been published. Time dependence has been investigated by e.g. /37/, /44/ and models in three space dimensions have been announced by e.g. /8/, /11/, /67/, /68/.

*Mathematics Research Center, University of Wisconsin, Madison, WI 53705, USA.

**Abteilung Physikalische Elecktronik, Institut fuer Allgemeine Elektrotechnik und Elektronik, TU Wien, Guszhausstrasze 27, A-1040, Wien, AUSTRIA.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2.

In spite of all these important and successful activities, the need for economic and highly user oriented computer programs became more and more apparent in the field of device modeling. Especially for MOS devices which have evolved since their invention by Kahng and Atalla /32/ to an incredible standard, modeling in two space dimensions has become inherently important because current flow controlled by a perpendicular field is an intrinsically two dimensional problem. One such program which has been applied successfully in many laboratories is called CADDET /59/. We have also tried to bridge that gap and developed MINIMOS /53/, /51/ for the two dimensional static analysis of planar MOS transistors.

2. Analysis of the Static Semiconductor Equations

In this chapter we review some of the existing analytical results for the fundamental semiconductor equations concerning existence and structure of their solutions. These results are of importance in both the theoretical and practical context, since - as we will see in the next chapter - the knowledge of the structure and smoothness properties of solutions is indeed essential for the development of a numerical solution method. The most familiar model of carrier transport in a semiconductor device has been proposed by Van Roosbroeck /61/. It consists of Poisson's equation (2.1), the current continuity equations for electrons (2.2) and holes (2.3) and the current relations for electrons (2.4) and holes (2.5)

$$\operatorname{div} \varepsilon \cdot \operatorname{grad} \psi = -q \cdot (p - n + C) \qquad (2.1)$$

$$\operatorname{div} \vec{J}_n = -q \cdot R \qquad (2.2)$$

$$\operatorname{div} \vec{J}_p = q \cdot R \qquad (2.3)$$

$$\vec{J}_n = -q \cdot (\mu_n \cdot n \cdot \operatorname{grad} \psi - D_n \cdot \operatorname{grad} n) \qquad (2.4)$$

$$\vec{J}_p = -q \cdot (\mu_p \cdot p \cdot \operatorname{grad} \psi + D_p \cdot \operatorname{grad} p) \qquad (2.5)$$

These relations form a system of coupled partial differential equations. Poisson's equation, coming from Maxwell's laws, describes the charge distribution in the interior of a semiconductor device. The balance of sinks and sources for electron- and hole currents is characterized by the continuity equations. The current relations describe the absolute value, direction and orientation of electron- and hole currents. The continuity equations and the current relations can be derived from Boltzmann's equation by not at all trivial means. It is not our intention to present in this paper the ideas behind these considerations. The interested reader is refered to /61/ and its secondary literature or text books on semiconductor physics e.g. /7/, /31/, /52/, /56/.

2.1 The Validity of the Basic Semiconductor Equations

It is of prime importance to be aware that equations (2.4) and (2.5) are not capable to describe exactly all phenomena occuring in real devices. For instance, they do not characterize effects which are caused by degenerate semiconductors (e.g. heavy doping). /38/, /60/, /63/ discuss some modifications of the current relations, which partially take into account the consequences introduced by degenerate semiconductors (e.g. invalidity of Boltzmann's statistics, bandgap narrowing). These modifications are not at all simple and lead to problems especially in the formulation of boundary conditions /47/, /62/. In case of modeling MOS devices, degeneracy, owing to the relatively low doping in the channel region, is practically irrelevant. For modern bipolar devices, though, bearing in mind shallow and extraordinarily heavily doped emitters, it is an absolute necessity to account for local degeneracy of the semiconductor.

Just as further examples (2.4) and (2.5) do not describe velocity overshoot phenomena which become apparent at feature lengths of 0.1μm for silicon and 1μm for gallium-arsenide /25/.

Certainly no effects which are due to ballistic transport (the existence of which is still questionable /30/) are included. The latter start to become important for feature sizes below 0.01μm for silicon and 0.1μm for gallium-arsenide /26/. Considering the state of the art of device miniaturization, neither effect has to bother the modelists of silicon devices. For gallium-arsenide devices new ideas are mandatory in the near future /25/, /46/, /45/.

2.2 Domain and Boundary Conditions

Most of the existing programs which solve the semiconductor equations are restricted to a rectangular device geometry. This is not essential as far as the analysis of the equations is concerned. In this chapter we shall assume that the equations (2.1)-(2.5) are posed in a domain D of R^n (n=1,2,3) with a piecewise smooth boundary ∂D. Equations (2.1)-(2.5) are subject to a mixed set of Dirichlet and Neumann boundary conditions. That means ∂D consists of three parts $\partial D = \partial D_1 \cup \partial D_2 \cup \partial D_3$. ∂D_1 denotes the part of the boundary where the device is surrounded by insulating material. There one assumes the boundary conditions:

$$\partial\psi/\partial\vec{n}_\perp = \partial n/\partial\vec{n}_\perp = \partial p/\partial\vec{n}_\perp = 0 \qquad (2.6)$$

Here $\vec{n}_\perp$ denotes the unit normal vector on ∂D which exists anywhere except at a finite number of points (arbitrarily defined corners of the simulation geometry). ∂D_2 denotes the part of the boundary corresponding to the ohmic contacts. There ψ, n and p are prescribed. The boundary conditions can be derived from the applied bias ψ_D and the assumptions of thermal equilibrium and vanishing space charge:

$$\psi = \psi_D + \psi_{\text{built-in}}, \quad n\cdot p = n_i^2, \quad n - p - C = 0 \qquad (2.7)$$

The last two conditions in (2.7) can be rewritten as:

$$n = (\sqrt{C^2+4\cdot n_i^2} + C)/2$$
$$p = (\sqrt{C^2+4\cdot n_i^2} - C)/2 \qquad (2.8)$$

In many applications it is desired to consider controlled insulator-semiconductor interfaces (e.g. MOS devices). So ∂D_3 denotes the part of the boundary which corresponds to such an interface. There we have the interface conditions:

$$\vec{J}_n\cdot\vec{n}| = \vec{J}_p\cdot\vec{n}| = 0$$
$$\varepsilon_{sem}\cdot\partial\psi/\partial\vec{n}_\perp|_{sem} = \varepsilon_{ins}\cdot\partial\psi/\partial\vec{n}_\perp|_{ins} \qquad (2.9)$$

Again $\vec{n}_\perp$ denotes the normal vector on ∂D. ε_{sem} and ε_{ins} denote the permittivity constants for the semiconductor and the insulator respectively. $\partial\psi/\partial\vec{n}_\perp|_{sem}$ and $\partial\psi/\partial\vec{n}_\perp|_{ins}$ denote the onesided limits of the derivatives perpendicular to the interface approaching the interface. Within the insulator the Laplace equation: div grad $\psi = 0$ holds.

2.3 Dependent Variables

For analytical purposes it is often useful to use other variables than n and p to describe the system (2.1)-(2.5). Two other sets of variables which are frequently employed are $(\psi,\varphi_n,\varphi_p)$ and (ψ,u,v) which relate to the set (ψ,n,p) by:

$$n = n_i\cdot e^{(\psi-\varphi_n)/U_t}, \quad p = n_i\cdot e^{(\varphi_p-\psi)/U_t} \qquad (2.10)$$

$$n = n_i\cdot e^{\psi/U_t}\cdot u, \quad p = n_i\cdot e^{-\psi/U_t}\cdot v \qquad (2.11)$$

(2.10) can be physically interpreted as the application of Boltzmann statistics. However (2.10) also can be regarded as a purely mathematical change of variables so that the question of the validity of the Boltzmann statistics does not need to be considered. The use of $(\psi,\varphi_n,\varphi_p)$ a priori excludes negative carrier densities n and p, which may be present as undesired nonphysical solutions of (2.1)-(2.5) if we use (ψ,n,p) or (ψ,u,v) as dependent variables. As we will see later in this chapter the advantage of the set (ψ,u,v) is that the continuity equations (2.2), (2.3) and current relations (2.4), (2.5) become self-adjoint. This also has an important impact on the use of iterative schemes for the solution of the evolving linear systems (cf. chapter 4). However, owing to the enormous range of the values of u and v, the sets (ψ,n,p) or $(\psi,\varphi_n,\varphi_p)$ have to be prefered for actual computations. We personally favour the set (ψ,n,p).

2.4 The Existence of Solutions and Scaling

The basic answer to the question of existence of solutions can be found in Mock /43/ or under slightly different assumptions in Bank, Jerome and Rose /5/. Both proofs are based on Schauder's fixpoint theorem. They are both valid for arbitrarily shaped domains and boundary conditions of the type previously described without an interface $(\partial D_3=\{\})$. Both papers consider the case of vanishing generation/recombination rate (R=0 in (2.2), (2.3)). In the setting of Mock (ψ,u,v) is used as dependent variables. The equations are scaled so that the intrinsic carrier density n_i, the thermal voltage U_t and the ratio elementary charge/permittivity are equal to unity. Thus, combining the continuity equations (2.2), (2.3) and current relations (2.4), (2.5), we have the system:

$$\operatorname{div}\operatorname{grad}\psi = e^{\psi}\cdot u - e^{-\psi}\cdot v - c \qquad (2.12)$$

$$\operatorname{div}(e^{\psi}\cdot\operatorname{grad} u) = 0 \qquad (2.13)$$

$$\operatorname{div}(e^{-\psi}\cdot\operatorname{grad} v) = 0 \qquad (2.14)$$

Then a map $M:\psi \rightarrow y$ is defined (details in /43/ or /4/) such that the evaluation of M requires the solution of (2.13) and (2.14) and a fixpoint ψ^* of M $(M(\psi^*)=\psi^*)$ together with the according functions (u,v) is a solution of the whole system (2.12)-(2.14). The existence of a fixpoint is shown by Schauder's fixpoint theorem. Questions concerning the degree of smoothness of these solutions (the existence of derivatives) are discussed in /42/.

However, Schauder's theorem is not constructive and does not indicate that iterating the map M will actually lead to the fixpoint. Moreover, it does not give any information about the structure of the solution which is of vital interest for actual computations. Since the dependent variables in the system (2.1)-(2.5) are of different order of magnitude and show a strongly different behaviour in regions with small and large space charge the first step towards a structural analysis of (2.1)-(2.5) has to be an appropriate scaling. A standard way of scaling (2.1)-(2.5) has been given by De Mari /14/. There ψ is scaled by the thermal voltage U_t, n and p are scaled by n_i (similar to Mock /43/) and the independent variables are scaled such that all multipying constants in Poisson's equation become unity. Although physically reasonable this approach has the disadvantage that n and p in general are still several orders of magnitude larger than ψ. A scaling which reduces ψ, n and p to the same order of magnitude has been given by Vasiliev'a and Butuzov /65/. This approach makes the system (2.1)-(2.5) accessible to an asymptotic analysis which is given together with applications in /40/, /41/ and /39/. There n and p are scaled by the maximum absolute value of the net doping C and the independent variables are scaled by the characteristic length of the device. More precisely the following scaling factors are employed.

quantity	symbol	value	
$\vec{x}$	l	$\max(\vec{x}-\vec{y})$, $\vec{x},\vec{y}$ in D	
ψ	U_t	$k \cdot T/q$	(2.15)
n,p	α	$\max\lvert C\rvert$	

After scaling the equations become:

$$\lambda^2 \cdot \text{div grad } \psi = n - p - C \tag{2.16}$$

$$\text{div (grad } n - n \cdot \text{grad } \psi \text{) } = -R$$

$$\text{div (grad } p + p \cdot \text{grad } \psi \text{) } = -R$$

Here, for simplicity only, μ_n and μ_p have been assumed to be constant. It should be noted that the following analysis also holds if the usual smooth dependence of μ_n and μ_p on n, p and grad ψ e.g./54/ is assumed. Since the independent variable x has been scaled, equations (2.16) are now posed on a domain D^s with maximal diameter equal to one. The small constant λ^2 multiplying the Laplacian in (2.16) is the minimal Debye length of the device:

$$\lambda^2 = \frac{\varepsilon \cdot Ut}{l^2 \cdot q \cdot \alpha} \tag{2.17}$$

l and α are defined in (2.15). Thus for high doping (α>>1) λ^2 will be small. For instance for a silicon device with characteristic length 25μm and $\alpha=10^{20}\text{cm}^{-3}$ we compute for λ^2 at approximate room temperature T=300K: $\lambda^2=4.10^{-10}$.

R denotes again the scaled generation/recombination rate. In the analysis given in /41/ the usual Shockley-Read-Hall term has been used which after scaling is of the form:

$$R = \frac{n \cdot p - (\gamma\lambda)^4}{n + p + 2 \cdot (\gamma\lambda)^2}, \quad \gamma = 1/2 \tag{2.18}$$

R is in general a (not necessarily mildly) nonlinear function of n,p and gradψ. Thus different models of R may influence the analytical results quite drastically. This is obviously to be expected as in many operating conditions the device behaviour depends strongly on the net generation/recombination R.

2.5 The Singular Perturbation Approach

(2.16) represents a singularly perturbed elliptic system with perturbation parameter λ. The advantage of this interpretation is that we can now obtain information about the structure of solutions of (2.16) by using asymptotic expansions: In the subdomains of D^S where the solutions behave smoothly we expand them into power series of the form:

$$w(\vec{x},\lambda) = \sum_{i=0}^{\infty} w_i^{\sim}(\vec{x}) \cdot \lambda^i, \quad w=(\psi,n,p)^T \tag{2.19}$$

which implies a smooth dependence on λ. C - the scaled doping - is smooth in these subdomains and exhibits a sharp transition across the pn-junctions in the device. For the case of an abrupt junction this behaviour is represented by a discontinuity across an n-1 dimensional manifold Γ:(x=x(s), s of R^{n-1}) in the device. Thus Γ is a point in 1 dimension, a curve in 2 dimensions and a surface in 3 dimensions. Of course one curve or surface has to be used for each junction. Since the procedure is the same for each of the junctions it is demonstrated only for one junction. In the case of an exponentially graded doping profile C consists of two parts:

$$C = C^{\sim} + C^{\wedge} \tag{2.20}$$

where $C^{\sim}$and $C^{\wedge}$ are discontinous, $C^{\sim}$ is piecewise smooth and $C^{\wedge}$ is exponentially decaying to zero away from Γ. In the vicinity of Γ the expansion (2.19) is not valid and has to be supplemented by a "layer" term acording to the singular perturbation analysis:

$$w(\vec{x},\lambda) = \sum_{i=0}^{\infty} [w_i^{\sim}(\vec{x}) + w_i^{\wedge}(s,t/\lambda)] \cdot \lambda^i, \quad w=(\psi,n,p)^T \tag{2.21}$$

Here the following coordinate transformation has been employed: For a point in the vicinity of Γ s denotes the parameter value at the nearest point on Γ and t denotes its distance perpendicular to Γ (cf. Fig. 1). Thus the solution of the semiconductor equations exhibits internal layers at pn-junctions.

The w_i^- and $w_i^\wedge$ in (2.21) can now be determined separately and the structure of the solution is given by its partition into the smooth part $\Sigma w_i^- \cdot \lambda^i$ and its rapidly varying part $\Sigma w_i^\wedge \cdot \lambda^i$. w_o^- has to satisfy the reduced equations:

$$0 = n_o^- - p_o^- - C^- \qquad (2.22)$$

$$\text{div } (\text{grad } n_o^- - n_o^- \cdot \text{grad}\psi_o^-) = -R^- \qquad (2.23)$$

$$\text{div } (\text{grad } p_o^- + p_o^- \cdot \text{grad}\psi_o^-) = -R^- \qquad (2.24)$$

For the sake of simplicity but without loss of generality the mobilities μ_n and μ_p have been assumed to be constant. (2.22)-(2.24) is subject to the boundary conditions (2.6)-(2.9). Of course the condition of vanishing space charge is redundant with (2.22). Since C^- is discontinous at Γ and (2.22)-(2.24) represents a second order system of two equations four "interface conditions" have to be imposed at Γ. They are of the form:

$$n_o^- \cdot e^{-\psi_o^-}|_{\vec{x}=\vec{x}-} = n_o^- \cdot e^{-\psi_o^-}|_{\vec{x}=\vec{x}+} \qquad (2.25)$$

$$p_o^- \cdot e^{\psi_o^-}|_{\vec{x}=\vec{x}-} = p_o^- \cdot e^{\psi_o^-}|_{\vec{x}=\vec{x}+} \qquad (2.26)$$

$$\vec{J}^-_{n_o} \cdot \vec{n}_\perp|_{\vec{x}=\vec{x}-} = \vec{J}^-_{n_o} \cdot \vec{n}_\perp|_{\vec{x}=\vec{x}+} \qquad (2.27)$$

$$\vec{J}^-_{p_o} \cdot \vec{n}_\perp|_{\vec{x}=\vec{x}-} = \vec{J}^-_{p_o} \cdot \vec{n}_\perp|_{\vec{x}=\vec{x}+} \qquad (2.28)$$

where $w|_{\vec{x}-}$ and $w|_{\vec{x}+}$ denote the onesided limits of w as $\vec{x}$ tends to Γ from each side. $\vec{n}_\perp$ denotes the unit normal vector on Γ. $\vec{J}^-_{n_o}$ and $\vec{J}^-_{p_o}$ are the zeroth order terms of the smooth parts of the (scaled) electron and hole current densities.

$$\vec{J}^{-}_{n_o} = \text{grad } n_o^- - n_o^- \cdot \text{grad } \psi_o^-$$
$$\vec{J}^{-}_{p_o} = \text{grad } p_o^- + p_o^- \cdot \text{grad } \psi_o^- \qquad (2.29)$$

(2.22)-(2.24) together with (2.25)-(2.28) and the boundary conditions (2.6)-(2.9) define the reduced problem whose solution is an $O(\lambda)$ approximation to the full solution away from Γ. As we will see in the next chapter the reduced problem is a useful tool for the development and analysis of numerical methods, since it (especially the conditions (2.25)-(2.28)) has to be solved implicitly by any discretisation method which requires a reasonable number of grid points.

The equations for the rapidly varying parts $\hat{w_i}$ reduce to ordinary differential equations. That means that only derivatives with respect to the "fast" variable t/λ occur. Since the rate of decay of $\hat{w_i}$ depends heavily on ψ the width of the layer grows with the applied voltage; a fact which is absolutely well known by device physicists, but which becomes nicely apparent by the singular perturbation approach.

3. Numerical Solution of the Semiconductor Equations

In this chapter we discuss some of the problems occuring in the numerical solution of the semiconductor equations and the analysis of existing numerical methods. From the viewpoint of numerical analysis there are essentially four major topics to be considered. The first one is the type of discretisation to be used. There exist programs for both Finite Element and Finite Difference discretisations of the system (2.1)-(2.5). As outlined in the previous chapter the solution exhibits a smooth behaviour in some subregions of the domain whereas in others it varies rapidly. Thus a nonuniform mesh is mandatory and adaptive mesh refinement is desirable. So the second topic is the question how to set up the mesh refinement algorithm i.e. which quantities have to be used to control the mesh. Each type of discretisation will lead to a large sparse system of nonlinear equations and so the solution of this system is the third topic. As fourth topic we discuss linear equations solvers which have to be used in topic three. For topics one to three many methods have been designed especially for the semiconductor equations. These points will be discussed in this chapter. For topic four standard numerical analysis is commonly used and so its discussion will be deferred to chapter four. For the sake of simplicity in nomenclature we shall only consider the two-dimensional case in this chapter. However, all results given in the following can be generalized to three dimensions in a straightforward manner. So, the equations are posed in a domain D of $\mathbb{R}^2$ and $\vec{x} = (x,y)^T$ denotes the independent variable.

3.1 Discretisation Schemes

Using Finite Elements or Finite Differences one has to take into account that Poisson's equation (2.1) is of a different type than the continuity equations. Poisson's equation - in the scaling of Markowich /40/ using the variables (ψ,u,v)

$$\lambda^2\cdot\text{div grad }\psi = e^{\psi}\cdot u - e^{-\psi}\cdot v - C \tag{3.1}$$

is a singularly perturbed elliptic problem whose right hand side has a positive derivative with respect to ψ. Thus it is of a standard form (as discussed in e.g. /22/) except for the discontinous or exponentially graded term C. Equations of that type are generally well behaved and it suffices to apply a usual discretisation scheme. In the case of Finite Differences equation (3.1) is discretized by:

$$\lambda^2\cdot(\text{div grad}_h\psi)_{ij} = n_{ij} - p_{ij} - C(x_i,y_j) \tag{3.2}$$

$$E^x_{i+1/2,j} = (\psi_{i+1,j}-\psi_{i,j})/h_i \qquad (3.3)$$

$$E^y_{i,j+1/2} = (\psi_{i,j+1}-\psi_{i,j})/k_j$$

$$h_i = x_{i+1}-x_i, \quad k_j = y_{j+1}-y_j$$

$$(\text{div grad } \psi)_{i,j} = 2\cdot(E^x_{i+1/2,j} - E^x_{i-1/2,j})/(h_i+h_{i-1}) + $$
$$+ 2\cdot(E^y_{i,j+1/2} - E^y_{i,j-1/2})/(k_j+k_{j-1}) \qquad (3.4)$$

Here ψ_{ij}, n_{ij} and p_{ij} denote the approximations to ψ, n and p at the gridpoint (x_i,y_j). $E^x_{i+1/2,j}$ denotes the value of $\partial\psi/\partial x$ at $(x_{i+1/2}=(x_i+x_{i+1})/2,\ y_j)$. $E^y_{i,j+1/2}$ denotes the value of $\partial\psi/\partial y$ at $(x_i,\ y_{j+1/2}=(y_j+y_{j+1})/2)$. If one of the neighbouring gridpoints (x_{i+1},y_j), (x_{i-1},y_j), (x_i,y_{j+1}), (x_i,y_{j-1}) does not exist - as possible in a terminating line approach /1/, /2/ or in the Finite Boxes approach /24/ - (3.4) has to be modified. We will go into some detail concerning these modifications in the next section. In the case of Finite Elements classical shape functions can be used (i.e. linear shape functions for triangular elements, bilinear shape functions for rectangular elements).

It turns out that the discretisation of the continuity equations is more crucial than the discretisation of Poissons's equation. The usual error analysis of discretisation methods provides an error estimate of the form:

$$\max |w_h-w| <= c\cdot H \qquad (3.5)$$

w_h denotes the numerical approximation to $w(x,y)=(\psi,n,p)^T$. H denotes the maximal gridspacing. The constant c will in general depend on the higher order derivatives of w. The singular perturbation analysis /41/ shows that derivatives of ψ^, n^ and p^ in (2.21) are of magnitude $O(\lambda^{-3})$ - $O(\lambda^{-4})$ locally near the junction (λ is defined in (2.17)). /41/ shows also that, even if a nonuniform mesh is used, the amount of gridpoints required to equidistribute the error term in (3.5) can be proportional to λ^{-2} which is of course prohibitive. Therefore a discretisation scheme is needed where the constant c in (3.5) does not depend on the higher derivatives of the rapidly varying terms $\hat{\psi}$, $\hat{n}$ and $\hat{p}$. For the case of Finite Differences such a scheme was given by Scharfetter and Gummel /50/. They approximate:

$$\vec{J}_n = \text{grad } n - n\cdot\text{grad } \psi \tag{3.6}$$

$$\text{div } \vec{J}_n = \partial J_n^x/\partial x + \partial J_n^y/\partial y = R \tag{3.7}$$

by:

$$\begin{aligned} J^x_{n\,i+1/2,j} &= \gamma((\psi_{i+1,j}-\psi_{i,j})/2)\cdot(n_{i+1,j}-n_{i,j})/h_i - \\ &\quad - (n_{i,j}+n_{i+1,j})/2\cdot(\psi_{i+1,j}-\psi_{i,j})/h_i \\ J^y_{n\,i,j+1/2} &= \gamma((\psi_{i,j+1}-\psi_{i,j})/2)\cdot(n_{i,j+1}-n_{i,j})/k_j - \\ &\quad - (n_{i,j}+n_{i,j+1})/2\cdot(\psi_{i,j+1}-\psi_{i,j})/k_j \\ \gamma(s) &= s\cdot\coth(s) \end{aligned} \tag{3.8}$$

$$\begin{aligned} &2\cdot(J^x_{n\,i+1/2,j} - J^x_{n\,i-1/2,j})/(h_i+h_{i-1}) + \\ &+ 2\cdot(J^y_{n\,i,j+1/2} - J^y_{n\,i,j-1/2})/(k_j+k_{j-1}) = R_{i,j} \end{aligned} \tag{3.9}$$

$J^x_{n\,i+1/2,j}$ denotes the value of J^x_n at $(x_{i+1/2}=(x_i+x_{i+1})/2,$ $y_j)$. $J^y_{n\,i,j+1/2}$ denotes the value of J^y_n at $(x_i,$ $y_{j+1/2}=(y_j+y_{j+1})/2)$. The continuity equation for holes is discretized analogously. Scharfetter and Gummel gave a physical reasoning for the derivation of their scheme. Markowich et al. /41/ proved that in one dimension the Scharfetter-Gummel scheme is uniformly convergent. That means that the error constant c in (3.5) does not depend on the derivatives of $\hat{\psi}$, $\hat{n}$ and $\hat{p}$ in (2.21) and therefore not on λ. For two dimensions /41/ shows that the choice $\gamma(s) = s\cdot\coth(s)$ is necessary for uniform convergence. Exponentially fitted schemes like the Scharfetter-Gummel scheme have been analyzed by Kellog /34/, /33/ and Doolan

/17/ (for different classes of problems). The reason for the uniform convergence of these schemes is that inside the pn-junction layers the interface conditions (2.25) and (2.26) are satisfied automatically if $|\mathrm{grad}\psi|$ is large and the gridspacing is not $O(\lambda)$.

The results for Finite Difference schemes suggest that a similiar approach (like the exponentially fitted schemes) should be used in the case of Finite Elements. This fact has been intuitively observed by Engel /21/ for the one-dimensional case. A modeling group at IBM has tried to make use of the Scharfetter-Gummel scheme for Finite Elements in two and three space dimensions /9/, /8/, /12/. However, we have the impression that their approach needs still quite a bit of analysis, although it has been used effectively by other modelists too e.g. /49/. Macheck /36/ has tried to develop a more rigorous discretisation for Finite Elements using exponentially fitted shape functions. He uses classical bilinear shape functions for ψ and

$$\begin{aligned}
\alpha_1(x,y) &= [1 - \varphi_1(x,y)]\cdot[1 - \varphi_2(x,y)] \\
\alpha_2(x,y) &= \varphi_1(x,y)\cdot[1 - \varphi_2(x,y)] \\
\alpha_3(x,y) &= \varphi_1(x,y)\cdot\varphi_2(x,y) \\
\alpha_4(x,y) &= [1 - \varphi_1(x,y)]\cdot\varphi_2(x,y)
\end{aligned}\tag{3.11}$$

for u, and

$$\begin{aligned}
\beta_1(x,y) &= [1 - \sigma_1(x,y)]\cdot[1 - \sigma_2(x,y)] \\
\beta_2(x,y) &= \sigma_1(x,y)\cdot[1 - \sigma_2(x,y)] \\
\beta_3(x,y) &= \sigma_1(x,y)\cdot\sigma_2(x,y) \\
\beta_4(x,y) &= [1 - \sigma_1(x,y)]\cdot\sigma_2(x,y)
\end{aligned}\tag{3.12}$$

for v, where

$$\Psi_1(x,y) = f(x,\frac{\partial\psi}{\partial x}) \qquad (3.13)$$

$$\Psi_2(x,y) = f(y,\frac{\partial\psi}{\partial y})$$

$$\sigma_1(x,y) = f(x,-\frac{\partial\psi}{\partial x})$$

$$\sigma_2(x,y) = f(y,-\frac{\partial\psi}{\partial y})$$

with: $f(x,a) = (\exp(ax)-1)/(\exp(a)-1)$ (3.14)

The advantage of these shape functions is that they accomodate nicely the layer behaviour of the solution. They degenerate into the ordinary bilinear shape functions when the electric potential is constant. In order to be able to switch from coarse to fine grid spacing in different subdomains transition elements have to be used (as outlined in the next section). However, no theoretical investigations have been carried out so far to analyse the uniform convergence properties of this method.

3.2 Grid Construction

Since subregions of strong variation of ψ, n and p alternate with regions where these quantities behave smoothly (i.e. their gradients are small) different meshsizes are mandatory in these subregions. Thus the discretisation scheme should be able to switch locally from a coarser to a finer grid. For the exponentially fitted (Scharfetter-Gummel) Finite Difference discretisation schemes this is done by the Finite Boxes approach /24/. Grid lines can terminate when the mesh is likely to be coarsened (cf. Fig.2). The point (x_{i+1},y_j) does not belong to the mesh. Thus the equations for the point (x_i,y_j) have to be modified since $\psi_{i+1,j}$, $n_{i+1,j}$ and $p_{i+1,j}$ are not available. This is done by proper interpolation between the (j-1)-st and (j+1)-st y-level. So (div grad $\psi)_{ij}$ is approximated by:

$$(\mathrm{div\ grad}\ \Psi)_{i,j} =$$

$$= 2\cdot((k_{j-1}\cdot E^{x}_{i+1/2,j+1} + k_j\cdot E^{x}_{i+1/2,j-1})/(k_j+k_{j-1}) -$$

$$- E^{x}_{i-1/2,j})/(h_i+h_{i-1}) +$$

$$+ 2\cdot(E^{y}_{i,j+1/2} - E^{y}_{i,j-1/2})/(k_j+k_{j-1}) \qquad (3.15)$$

$E^{x}_{i-1/2,j}$, $E^{y}_{i,j+1/2}$ etc. are defined in (3.3). The continuity equations are approximated by:

$$2\cdot((k_{j-1}\cdot J^{x}_{n\,i+1/2,j+1} + k_j\cdot J^{x}_{n\,i+1/2,j-1})/(k_j+k_{j-1}) -$$

$$- J^{x}_{n\,i-1/2,j})/(h_i+h_{i-1}) +$$

$$+ 2\cdot(J^{y}_{n\,i,j+1/2} - J^{y}_{n\,i,j-1/2})/(k_j+k_{j-1}) = R_{i,j} \qquad (3.16)$$

$J^{x}_{n\,i-1/2,j}$, $J^{y}_{n\,i,j+1/2}$ etc. are defined in (3.8). For reasons of numerical stability only one gridline is allowed to terminate at a box. This approach is a generalisation of the "Terminating Line" approach introduced by Adler /1/, /2/ as already mentioned.

In the Finite Element approach of Macheck /36/ transition elements composed of three triangles are used to coarsen the mesh locally (cf. Fig.3). Within these triangles a different set of shape functions has to be used. They are derived by holding the current densities $\vec{J}_n$ and $\vec{J}_p$ constant along the edges of a triangle similar to the approach of /10/.

In the Finite Element as well as in the Finite Difference (Boxes) approach the question arises which criteria should be used to generate the mesh. If the user of a simulation program has to define his elements or nodes a priori as input parameters, this could perhaps be done by experience /10/. However, if - as it is the case for modern user oriented programs - an adaptive mesh selection is desired mathematically formulated criteria are a "sine qua non". Generally such criteria should satisfy two conditions. Firstly they should not cause the program to construct more gridpoints/elements than necessary to achieve a certain accuracy. Secondly they should guarantee that a prescribed relative accuracy δ is really achieved once they are satisfied. A usual way to design adaptive mesh refinement procedures is to equidistribute the local truncation error of the discretisation scheme. In the case of Finite Differences this error is proportional to the meshsize and the third and fourth derivatives of ψ, n and p. Markowich /41/ however showed that it is practically not possible to equidistribute this quantity. In the case of a simple MOS-transistor $O(\delta^{-2}\lambda^{-2})$ gridpoints would be required. On the other hand the singular perturbation analysis shows that the solution of the difference scheme approximates the solution of the reduced problem (2.22)-(2.24) even if this criterion is not satisfied inside the layer regions (inversion layer and space charge regions). Therefore the quantity to be equidistributed is the discretisation error of Poisson's equation (i.e. the partial derivatives of the space charge times the meshsizes). This equidistribution can be relaxed inside the pn-junction layers by e.g. simply limiting the number of gridpoints there.

3.3 Linearisation Schemes

Each discretisation scheme (Finite Differences or Finite Elements) will lead to a large sparse system of nonlinear equations to be solved. The theory of iterative methods to solve these equations is to a large extent independent of the used discretisation and so it is convenient to view the whole problem as solving a nonlinear system of equations iteratively by solving linear systems. The existing numerical methods can essentially be divided into two classes: The first approach, a block nonlinear iteration algorithm, is due to Gummel /29/ and uses the fact that the current relations are linear in the variables u and v (as defined in (2.11)). In these variables the equations become (again we use the scaling of /36/):

$$\lambda^2 \cdot \mathrm{div\ grad}\ \psi = e^{\psi} \cdot u - e^{-\psi} \cdot v - C \tag{3.17}$$

$$\mathrm{div}\ \vec{J}_n = R, \quad \vec{J}_n = e^{\psi} \cdot \mathrm{grad}\ u \tag{3.18}$$

$$\mathrm{div}\ \vec{J}_p = -R, \quad \vec{J}_p = -e^{-\psi} \cdot \mathrm{grad}\ v \tag{3.19}$$

Gummels approach works as follows: Given $(\psi,u,v)^k$ ψ^{k+1} is computed by solving:

$$\lambda^2 \cdot \mathrm{div\ grad}\ \psi^{k+1} = e^{\psi^{k+1}} \cdot u^k - e^{-\psi^{k+1}} \cdot v^k - C \tag{3.20}$$

subject to the appropriate boundary conditions. Then u^{k+1} and v^{k+1} are computed from:

$$\mathrm{div}\ \vec{J}_n^{k+1} = R(\mathrm{grad}\ \psi^{k+1}, u^k, v^k), \quad \vec{J}_n^{k+1} = e^{\psi^{k+1}} \cdot \mathrm{grad}\ u^{k+1} \tag{3.21}$$

$$\mathrm{div}\ \vec{J}_p^{k+1} = -R(\mathrm{grad}\ \psi^{k+1}, u^k, v^k), \quad \vec{J}_p^{k+1} = -e^{-\psi^{k+1}} \cdot \mathrm{grad}\ v^{k+1} \tag{3.22}$$

together with the boundary conditions for u and v. (3.21) and (3.22) are two decoupled linear equations for u^{k+1} and v^{k+1}. Poissons's equation (3.20) is nonlinear in this setting and therefore it has to be solved iteratively itself in each step by a Newton like method. Since Newton's method is an inner iteration within the overall iteration process (3.20)-(3.22) it may not be necessary to let this inner iteration "fully converge" /27/. It could for instance be considered to do only one Newton step for each iteration. This would lead to the linear equation:

$$\lambda^2 \cdot \mathrm{div\ grad}\ \psi^{k+1} = (e^{\psi^k} \cdot u^k + e^{-\psi^k} \cdot v^k) \cdot (\psi^{k+1} - \psi^k) + \\ + e^{\psi^k} \cdot u^k - e^{-\psi^k} \cdot v^k - C \tag{3.23}$$

instead of (3.20). The advantage of Gummels's method is obvious. (3.20)-(3.22) can be solved sequentially which decreases the required amount of storage and computing time drastically for each step. However, bad convergence properties can be observed in the case of high currents. This is explained by viewing

(3.20)-(3.22) as iterating the map $M:(u^k,v^k)\rightarrow(u^{k+1},v^{k+1})$ where the evaluation of M involves the solution of (3.20). Then the norm of the linearisation of M (as an operator acting in the appropriate spaces) at the fixpoint $M(u^*,v^*)=(u^*,v^*)$ is proportional to the current densities /42/.

The second approach to the solution of the nonlinear equations (2.1)-(2.5) is a damped modified Newton method. To solve the general equation $F(x)=0$ one computes the sequence $<x^k>$ by:

$$M^k\cdot\delta^k = -F(x^k),\ x^{k+1} = x^k + t^k\cdot\delta^k \tag{3.24}$$

For the usual Newton method $M^k = F'(x^k)$ and $t^k = 1$ holds. Bank and Rose /4/ have given criteria for the choice of the damping parameters t^k which guarantee global convergence. Moreover they investigate how well δ^k has to approximate the classical Newton step in order to get a certain rate of convergence. They obtain that the rate of convergence is p $(1<p<2)$ if:

$$|M^k\cdot\delta^k + F(x^k)| = O(|F(x^k)|^p) \tag{3.25}$$

holds asymptotically for $k\rightarrow\infty$. Alternatively Bank and Rose /3/ suggested $M^k = \lambda^k I + F'(x^k)$ where λ^k is proportional to $|F(x^k)|$. Franz /24/ tested this method with good success. However, he additionally chooses damping parameters t^k according to Deuflhard /15/, /16/.

Since this approach has the disadvantage that all three equations are solved simultaneously - and therefore the storage requirements are fairly large - we suggest a Block-Newton-SOR method /24/. Defining $F=(F_1,F_2,F_3)^T$ Newton's method at step k is:

$$\begin{Bmatrix} \frac{\partial F_1}{\partial \psi} & \frac{\partial F_1}{\partial n} & \frac{\partial F_1}{\partial p} \\ \frac{\partial F_2}{\partial \psi} & \frac{\partial F_2}{\partial n} & \frac{\partial F_2}{\partial p} \\ \frac{\partial F_3}{\partial \psi} & \frac{\partial F_3}{\partial n} & \frac{\partial F_3}{\partial p} \end{Bmatrix}^k \cdot \begin{Bmatrix} \delta\psi^k \\ \delta n^k \\ \delta p^k \end{Bmatrix} = - \begin{Bmatrix} F_1(\psi^k,n^k,p^k) \\ F_2(\psi^k,n^k,p^k) \\ F_3(\psi^k,n^k,p^k) \end{Bmatrix} . \tag{3.26}$$

Under the assumption that the Jacobian is definite one can use a classical block iteration scheme (iteration index m) for the solution of the k-th Newton step:

$$\begin{Bmatrix} \frac{\partial F_1}{\partial \psi} & 0 & 0 \\ \frac{\partial F_2}{\partial \psi} & \frac{\partial F_2}{\partial n} & 0 \\ \frac{\partial F_3}{\partial \psi} & \frac{\partial F_3}{\partial n} & \frac{\partial F_3}{\partial p} \end{Bmatrix}^k \cdot \begin{Bmatrix} \delta\psi^k \\ \delta n^k \\ \delta p^k \end{Bmatrix}^{m+1} = - \begin{Bmatrix} F_1(\psi^k,n^k,p^k) \\ F_2(\psi^k,n^k,p^k) \\ F_3(\psi^k,n^k,p^k) \end{Bmatrix} - \begin{Bmatrix} 0 & \frac{\partial F_1}{\partial n} & \frac{\partial F_1}{\partial p} \\ 0 & 0 & \frac{\partial F_2}{\partial p} \\ 0 & 0 & 0 \end{Bmatrix}^k \begin{Bmatrix} \delta\psi^k \\ \delta n^k \\ \delta p^k \end{Bmatrix}^m \tag{3.27}$$

Since the coefficient matrix of (3.27) is block lower triangular one can decouple the elimination process into three linear systems (3.28)-(3.30) which have to be solved sequentially.

$$\frac{\partial F_1}{\partial \psi}^k \cdot \delta\psi^{km+1} = -F_1(\psi^k,n^k,p^k) - \frac{\partial F_1}{\partial n}^k \cdot \delta n^{km} - \frac{\partial F_1}{\partial p}^k \cdot \delta p^{km} \tag{3.28}$$

$$\frac{\partial F_2}{\partial n}^k \cdot \delta n^{km+1} = -F_2(\psi^k,n^k,p^k) - \frac{\partial F_2}{\partial \psi}^k \cdot \delta\psi^{km+1} - \frac{\partial F_2}{\partial p}^k \cdot \delta p^{km} \tag{3.29}$$

$$\frac{\partial F_3}{\partial p}^k \cdot \delta p^{km+1} = -F_3(\psi^k,n^k,p^k) - \frac{\partial F_3}{\partial \psi}^k \cdot \delta\psi^{km+1} - \frac{\partial F_3}{\partial n}^k \cdot \delta n^{km+1} \tag{3.30}$$

This iteration method has (like Gummel's method) the advantage that the equations can be solved sequentially. To end up with the Block-Newton-SOR method one has to resubstitute the series expansions on the right hand side of (3.28)-(3.30) and to introduce a relaxation parameter ω:

$$\frac{\partial F_1{}^k}{\partial \psi} \cdot \delta\psi^{km+1} = -\omega \cdot F_1(\psi^k, n^k+\delta n^{km}, p^k+\delta p^{km}) \tag{3.31}$$

$$\frac{\partial F_2{}^k}{\partial n} \cdot \delta n^{km+1} = -\omega \cdot F_2(\psi^k+\delta\psi^{km+1}, n^k, p^k+\delta p^{km}) \tag{3.32}$$

$$\frac{\partial F_3{}^k}{\partial p} \cdot \delta p^{km+1} = -\omega \cdot F_3(\psi^k+\delta\psi^{km+1}, n^k+\delta n^{km+1}, p^k) \tag{3.33}$$

This method converges linearly /48/. However, we still have to perform thorough investigations in order to properly judge the convergence properties.

4. Solution of Linear Systems

For any of the linearization procedures which have been outlined in the last chapter a large sparse linear equation system (4.1) has to be solved repeatedly.

$$A \cdot x = b \tag{4.1}$$

A has been derived by linearizing discretized PDEs. Hence A has only five to nine nonzero entries per row and block (the blocks are defined in (3.26)); A is very sparse. For the solution of these special types of linear systems of equations two classes of methods, can, in principle, be used: direct methods which are based on elimination and iterative methods. An excellent survey on that subject has been published recently by Duff /18/. Classical Gaussian elimination is not feasible for our systems of equations because the rank of A in (4.1) is very large and A has many coefficients which are zero. Therefore, modifications of the classical Gaussian elimination algorithm have to be introduced to account for the zero entries. There exist quite a few activities on that subject (c.f. /19/) and powerful algorithms which treat the nonzero coefficients only are available (so called sparse matrix codes). Another serious drawback of direct methods lies in the fact that the upper triangular matrix which is created by the elimination process has to be stored for back substitution. This matrix has usually more nonzero entries than the matrix A. Therefore, memory requirement of direct methods is substantial. One advantage of the linear systems obtained from the discretised semiconductor equations is that no pivoting in order to maintain numerical stability is needed. In spite of all drawbacks of direct methods, their major advantage is high accuracy of the solution. However, we feel that for the semiconductor problems iterative algorithms are to emphasize. Nevertheless we and many others have observed difficulties with respect to the convergence speed of iterative methods, so that the direct methods, which require an exactly predictable amount of computer resources, will always stay in consideration.

The fundamental idea of relaxation methods (which are the best established iterative methods) is the splitting of the coefficient matrix A (4.1) into three matrices D, E, F (4.2).

$$A = D - E - F \tag{4.2}$$

D denotes the diagonal entries of A; -E denotes a lower triangular matrix which consists of all sub-diagonal entries of A; and -F denotes an upper triangular matrix which consists of all super-diagonal entries of A.

With an arbitrary non singular matrix B which has the same rank as A the linear system (4.1) can be rewritten to (4.3):

$$B \cdot x + (A-B) \cdot x = b \tag{4.3}$$

One obtains an iterative scheme by setting:

$$B \cdot x^{k+1} = b - (A-B) \cdot x^{k} \tag{4.4}$$

(4.4) can be solved for x^{k+1}:

$$x^{k+1} = (I-B^{-1} \cdot A) \cdot x^{k} + B^{-1} \cdot b \tag{4.5}$$

The scheme (4.5) will converge if condition (4.6) holds:

$$\rho(I-B^{-1} \cdot A) < 1 \tag{4.6}$$

(4.6) is a necessary and sufficient condition where ρ denotes the spectral radius /64/. Any relaxation method can be derived by differently choosing the matrix B from the splitting of A (4.2). The simplest scheme, the point-Jacobi method, uses D for B. Matrix D is a diagonal matrix and, therefore, is easily invertible. The Gauss-Seidel method uses D-E for B. The matrix D-E is a lower triangular matrix. Therefore one has only to perform a forward substitution process for its inversion. The successive overrelaxation method (SOR) uses a parameter ω within the range]0,2[. The iteration matrix B is defined:

$$B = D/\omega - E \tag{4.7}$$

Since B is again a lower triangular matrix, its inversion is instantly reduced to a substitution.

The major advantage of these iterative methods lies in their simplicity. They are very easy to program and demand only low memory requirement. As already noted, they converge if condition (4.6) holds. However, this is generally difficult to prove. A sufficient condition for convergence is that A is positive definite (4.8) which is the normal case for five-point-star discretized PDEs.

$$x^{T} \cdot A \cdot x > 0 \text{ for all } x \neq 0 \tag{4.8}$$

It should be noted again here that the current relations and continuity equations are not self adjoint if (ψ,n,p) are used as variables (see (2.10), (2.11)). However, the transformation:

$$n = e^{\psi}\cdot u, \quad p = e^{-\psi}\cdot v \tag{4.9}$$

results in a similarity transformation of the iteration matrix in (4.6). Thus the spectral radius of the iteration matrix is not influenced and the same convergence properties are obtained as if the system had been discretized in its self adjoint form with (ψ,u,v) as variables.

Some point-iterative schemes can by accelerated quite remarkably with the conjugate gradient method or the Chebyshev method. An excellent survey on these topics can be found in /28/.

Various activities can be observed for the development of more powerful algorithms with the advantages of iterative schemes. One of the best known algorithms which has been established in semiconductor device analysis is Stone's strongly implicit procedure /58/. Stone's idea was to modify the original coefficient matrix A by adding a matrix N (whose norm is much smaller than the norm of A) so that a factorization of (A+N) involves less computational effort than the standard decomposition of A. Assuming this has been done, the development of an iterative procedure is then fairly straightforward because the equation can be written as:

$$(A+N)\cdot x = (A+N)\cdot x + (b-A\cdot x) \tag{4.10}$$

which suggests the iterative procedure:

$$(A+N)\cdot x^{k+1} = (A+N)\cdot x^{k} + (b-A\cdot x^{k}) \tag{4.11}$$

When the right hand side is known and if (A+N) can be factorized easily, (4.11) gives an efficient method for directly solving for x^{k+1}. Furthermore, one would intuitively expect a rapid rate of convergence if N is sufficiently small compared to

A. We will refrain from explaining in detail Stone's suggestion of how to choose the perturbation matrix N because this has been done thoroughly in many publications e.g. /23/, /55/, /58/. A major disadvantage of Stone's method is that it is only applicable for linear systems obtained by a classical Finite Difference discretisation. It is not applicable for systems obtained by the Finite Boxes approach or the general Finite Element approach.

There exist a few algorithms which are similar to Stone's method in terms of underlying ideas. The most attractive are the method of Dupont et al. /20/, the "alternating direction implicit" methods e.g. /6/, /23/, /66/ and the Fourier methods /57/, /64/. However, most of these sophisticated algorithms lack general applicability.

No matter which iterative method is used one has to deal with the question of an appropriate termination (convergence) criterion. Usually (4.12) is applied with a properly chosen relative accuracy ε:

$$|x^{k+1}-x^{k}| < \varepsilon \cdot |x^{k+1}| \qquad (4.12)$$

Since increments still accumulate when (4.12) is already satisfied we suggest to use (4.13) instead of (4.12):

$$|x^{k+1}-x^{k}| < \varepsilon \cdot |x^{k+1}| \cdot (1-\varrho(G)) \qquad (4.13)$$

$\varrho(G)$ can be estimated as $\lim\limits_{k \to \infty} |x^{k+1}-x^{k}|/|x^{k}-x^{k-1}|$.

One disadvantage of all strongly implicit methods and also the direct methods is that they cannot be implemented efficiently on a computer with a pipe-line architecture (vector processor). Some comments on that subject have been given in /18/.

5. A Glimpse on Results

As an illustrative example a relatively simple structure, a two dimensional diode, is chosen. Fig. 4 shows the doping profile as birds-eye-view plot. A substrate with $10^{14}cm^{-3}$ acceptor concentration and an exponentially graded n-region with $10^{19}cm^{-3}$ maximum doping is assumed. The initial mesh is automatically generated from the doping profile and the geometry definition. The simulation domain (device geometry) is a square of 100µ times 100µm size. At the n-region an ohmic contact with length 20µm is assumed. The substrate is fully contacted. The initial mesh for a Finite Boxes program is shown in Fig. 5 and for a Finite Element program in Fig. 6. The point allocation is identical for both representations. The grid consists of 121 points versus 178 when all gridlines are extended throughout the device. This clearly demonstrates the advantage of the Finite Boxes approach. In Finite Element representation one has to deal with 80 rectangular elements and 17 transition elements which consist of 51 triangles.

Fig. 7 shows the final grid for an operating condition of 0.7V forward bias in Finite Boxes representation. This mesh is obtained after several adaption processes using the criteria given in chapter 3. It consists of 270 points (versus 480 for the classical approach). In Fig. 8 the potential distribution is drawn. From this plot and even better from the electron density (Fig. 9) one nicely can deduce the effects of high injection. E.g. the substrate is flooded with carriers. Fig. 10 shows the magnitude of the electron current density. The peak value is about 180 A/cm^2. The sharply pronounced peak which exists at the transition of the Dirichlet boundary condition to the Neumann boundary condition corresponds to a singularity of the carrier densities. Physically interpreted this effect is well known as contact-corner-current-crowding.

Fig. 11 shows the final grid for an operating condition of -20V (reverse) bias in Finite Element representation. This mesh consists of 363 points (625 for classical Finite Differences) which correspond to 277 rectangular elements and 41 transition elements (123 triangles). The electron density for this operating point is given in Fig. 12. One nicely observes the depletion region and the typical shape of the drop of the electron density in that region owing to thermal generation. In Fig. 13 the magnitude of the electron current density is drawn. The singularity at the contact corner is, although it still exists, not so pronounced. Note that there are about seven orders of magnitude difference in the peak value compared to Fig. 10.

6. Conclusion

In this paper we have presented an analysis of the steady state semiconductor equations and the impact of this analysis on the design of device simulation programs. By appropriate scaling we have transformed the semiconductor equations into a singularly perturbed elliptic system with nonsmooth data. Information obtained from the singular perturbation analysis has been used to investigate stability and convergence of discretisation schemes with particular emphasis on the adaptive construction of efficient grids. We have reviewed algorithms for the solution of nonlinear and linear systems of the discretized semiconductor equations. An example has demonstrated the power and flexibility a device simulation program can achieve when using the information we have presented for program design.

Acknowledgement

This work is sponsored by the "Fonds zur Foerderung der wissenschaftlichen Forschung" Project No. S22/11, by the Research Lab. of Siemens AG Munich and by the United States Army under Contract No. DAAG-29-80-C-0041. Helpful discussions with Dr.P.Markowich and Prof.Dr.H.Poetzl who also critically read the manuscript were highly appreciated. We should also like to thank Dipl.Ing.A.Franz and Dipl.Ing.G.Franz for providing the computer plots.

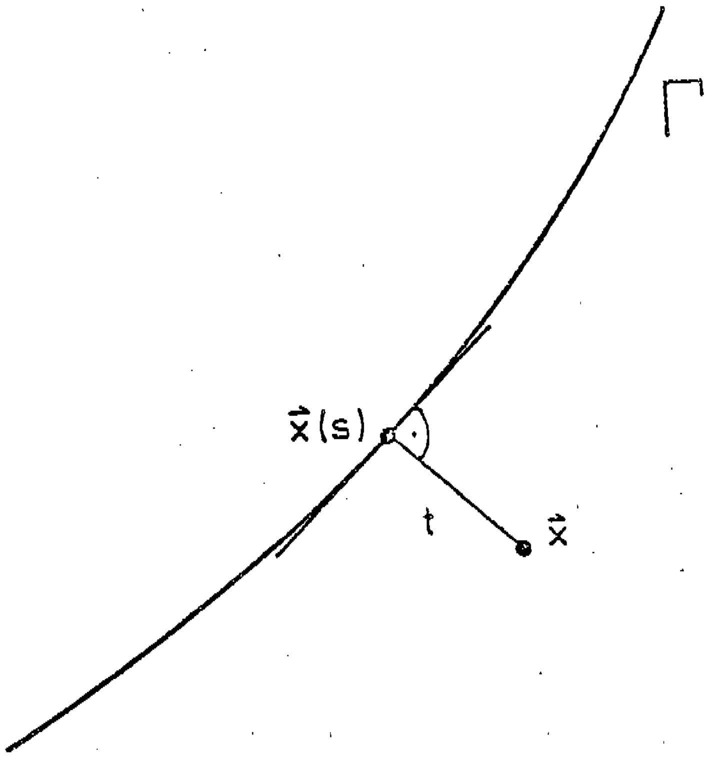


Fig. 1 Local Coordinates of the Layer Solution

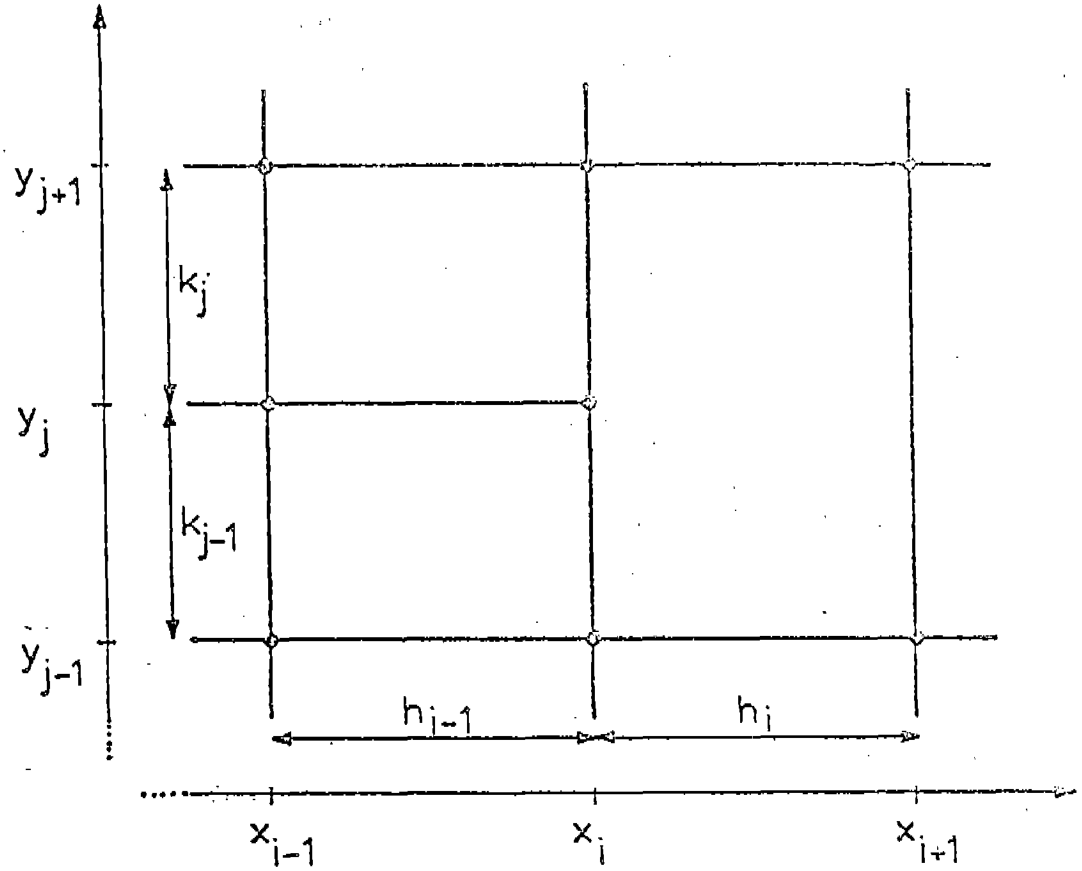


Fig. 2 A Typical Finite Boxes Configuration

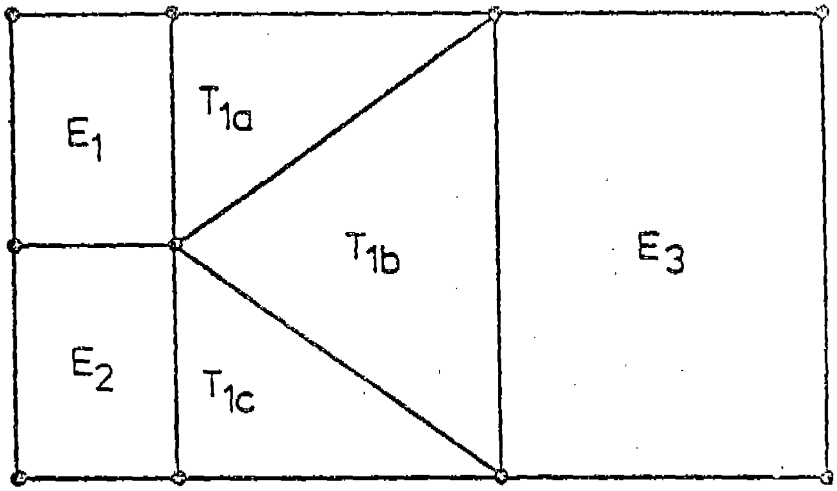


Fig. 3 A Transition Element to Coarsen a Mesh

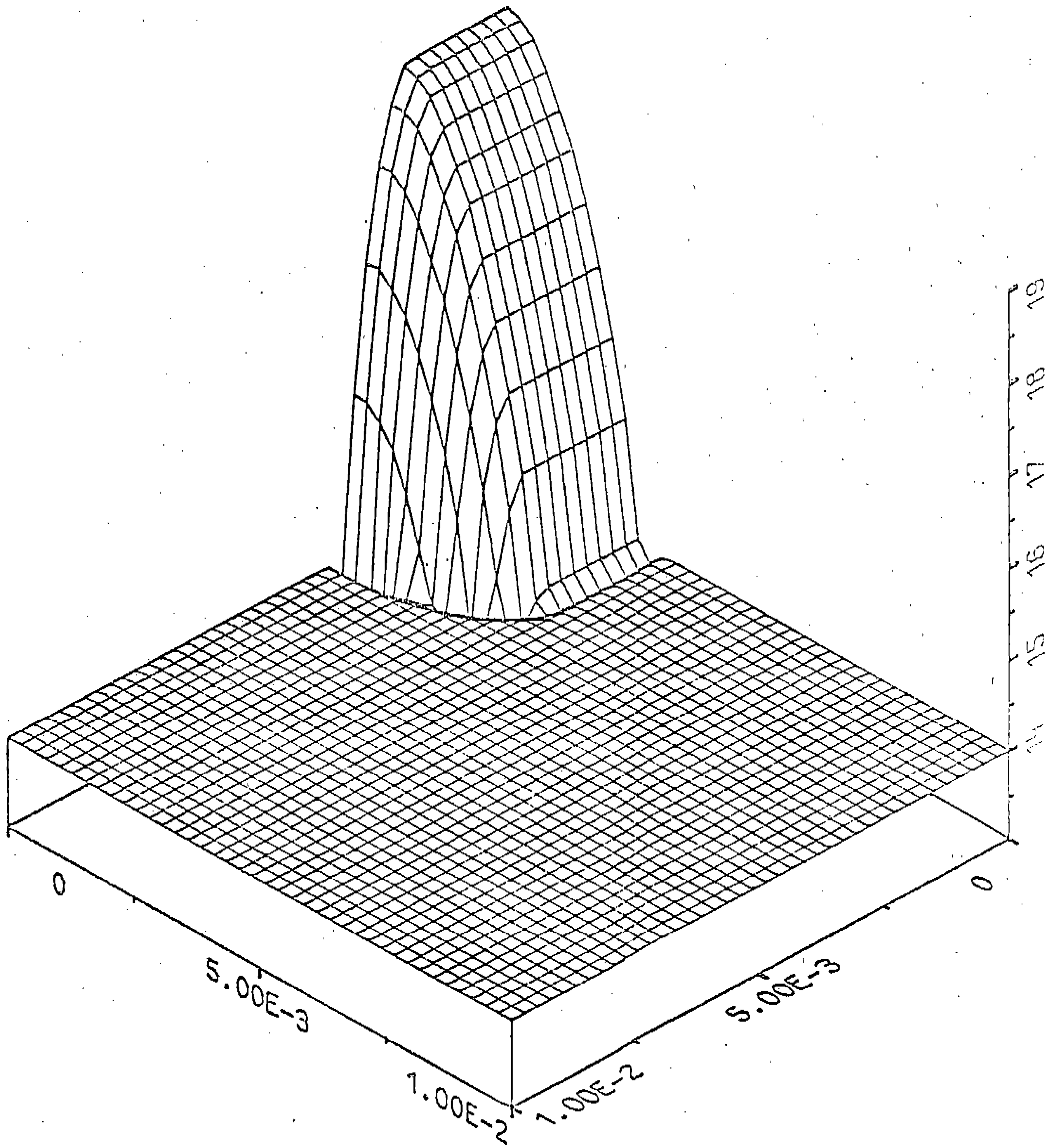


Fig. 4 Doping Profile $[cm^{-3}]$ (log.)

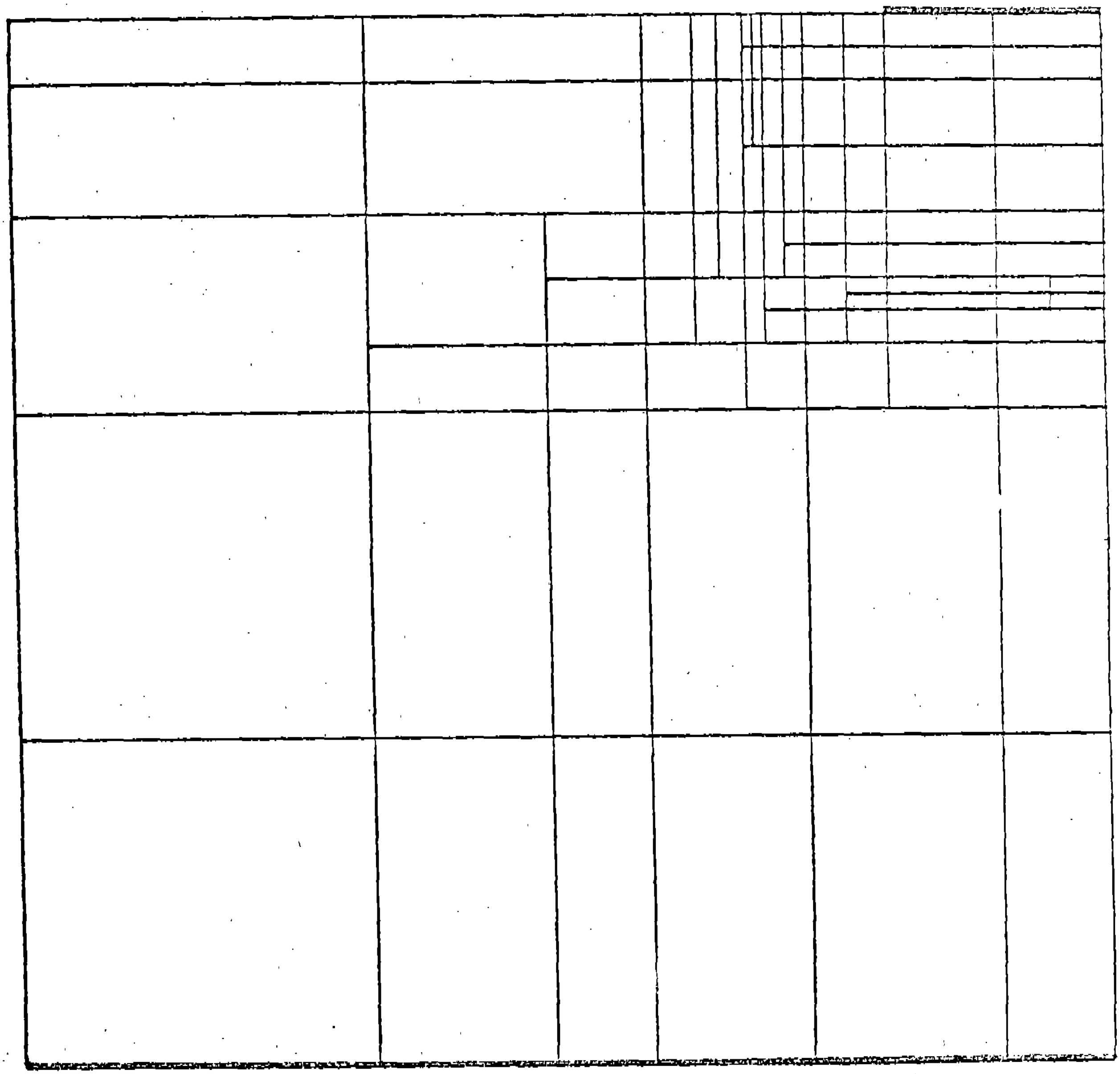


Fig. 5 Initial Mesh in Finite Boxes Interpretation

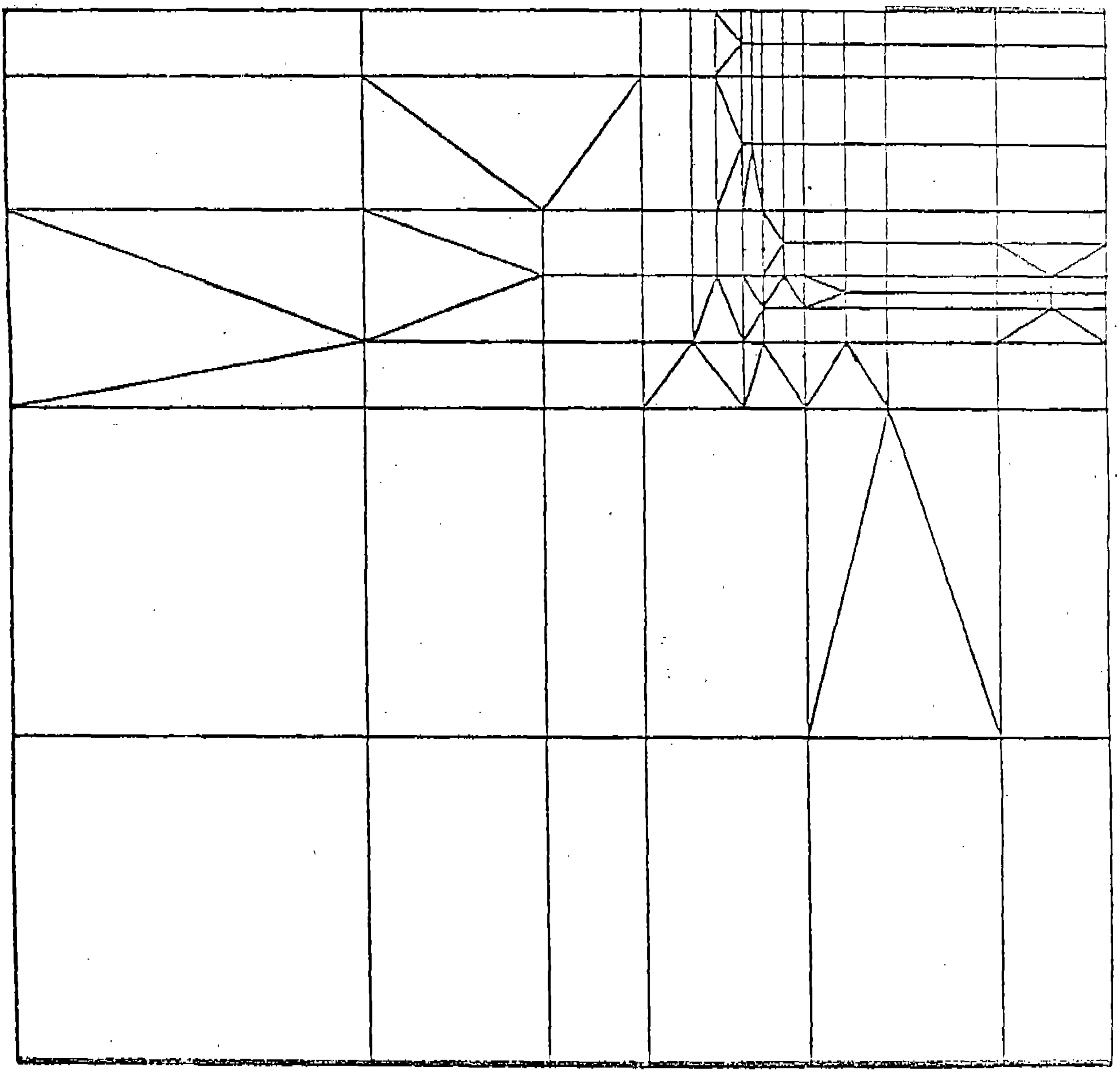


Fig. 6 Initial Mesh in Finite Element Interpretation

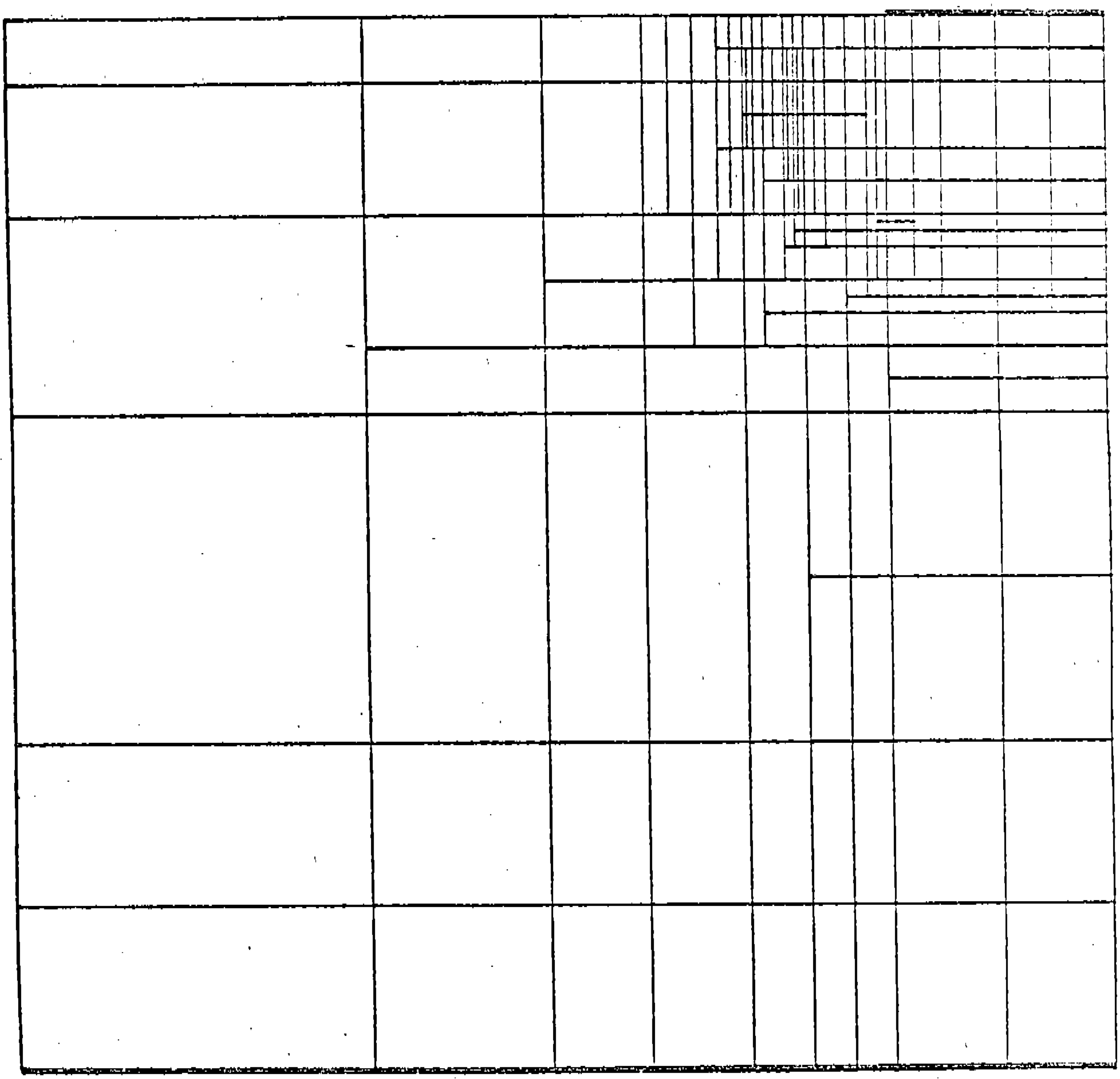


Fig. 7 Final Mesh for 0.7V Forward Bias (Finite Boxes)

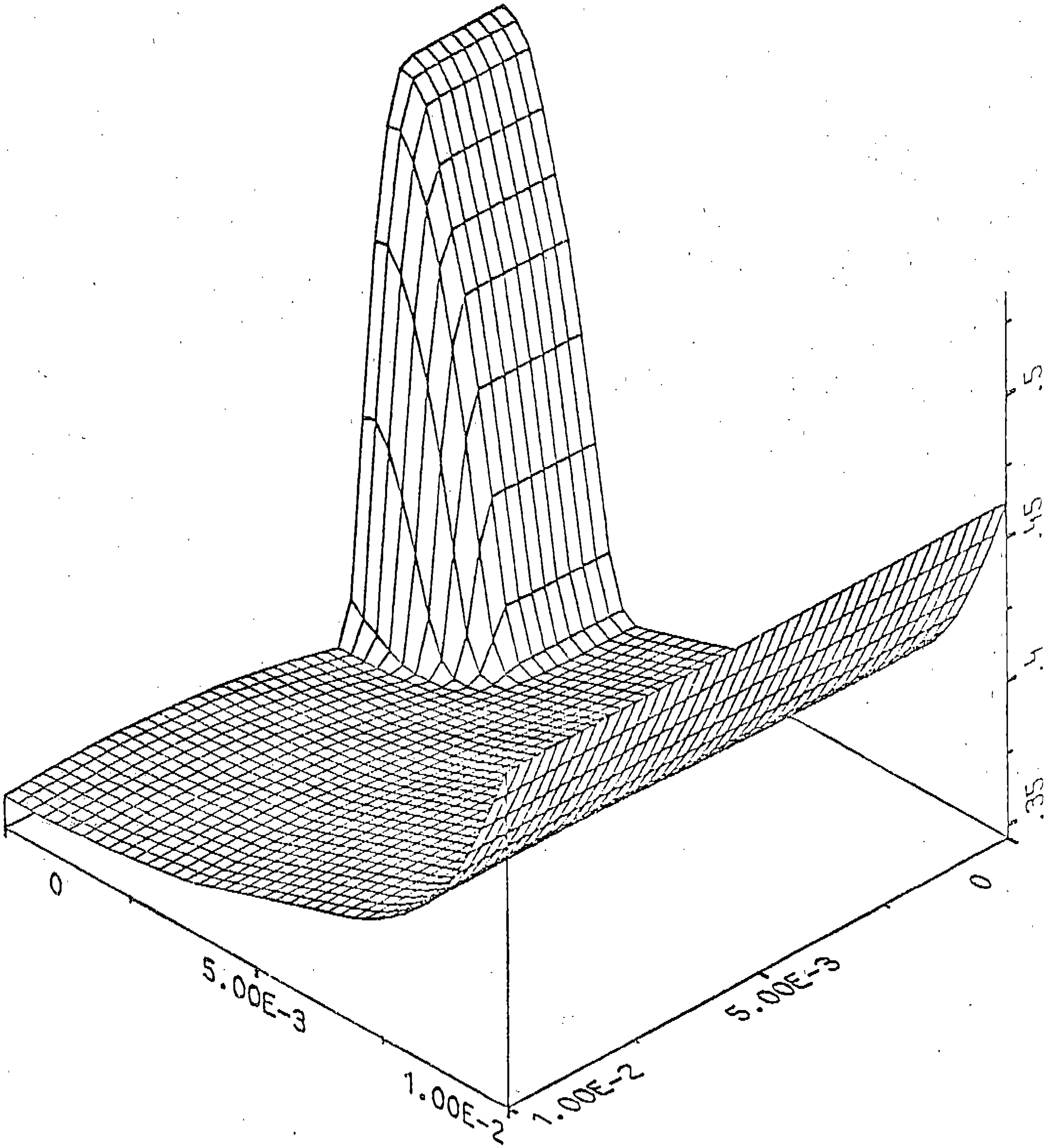


Fig. 8 Potential distribution (0.7V) [V] (lin)

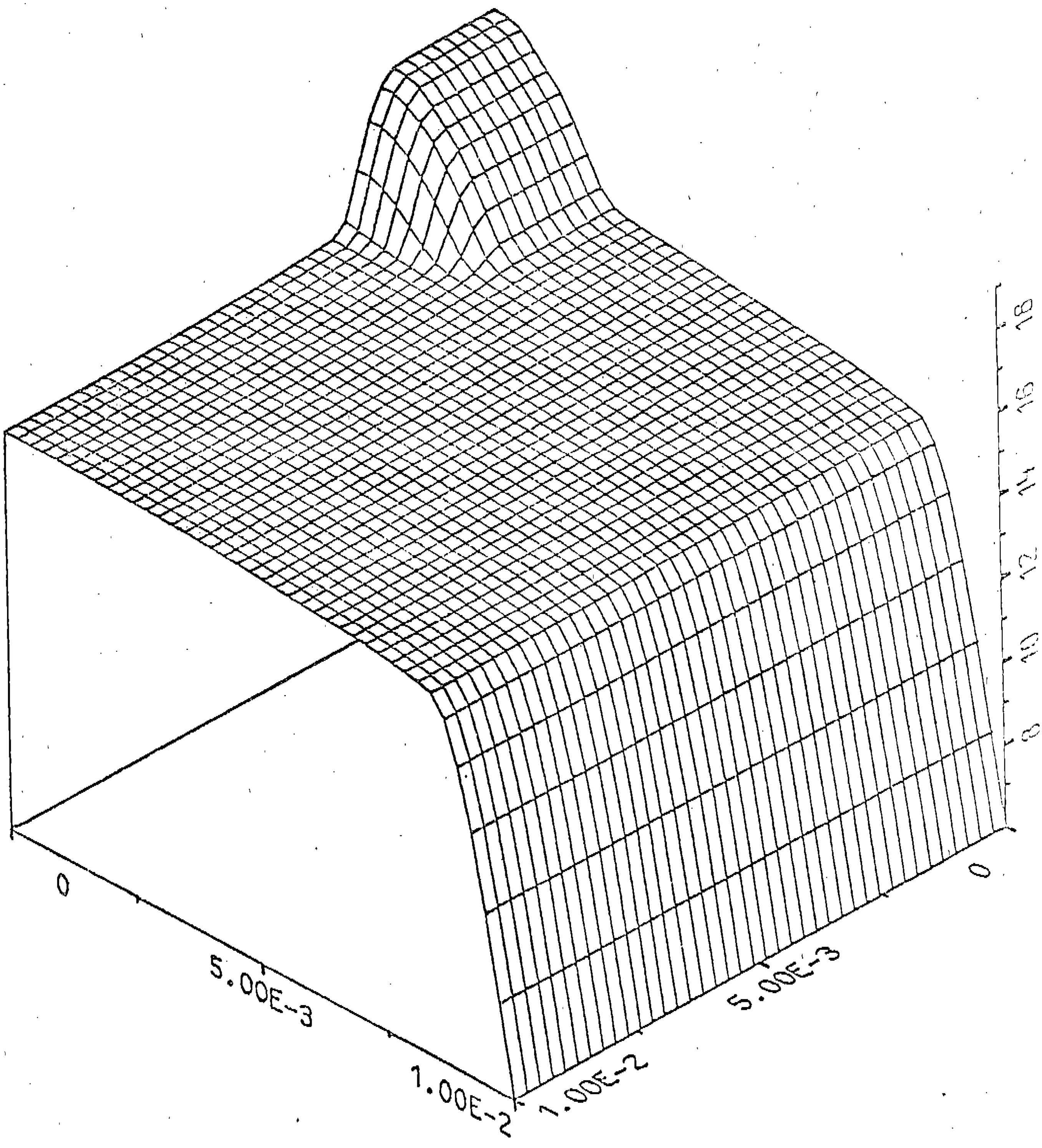


Fig. 9 Electron concentration (0.7V) [cm^{-3}] (log)

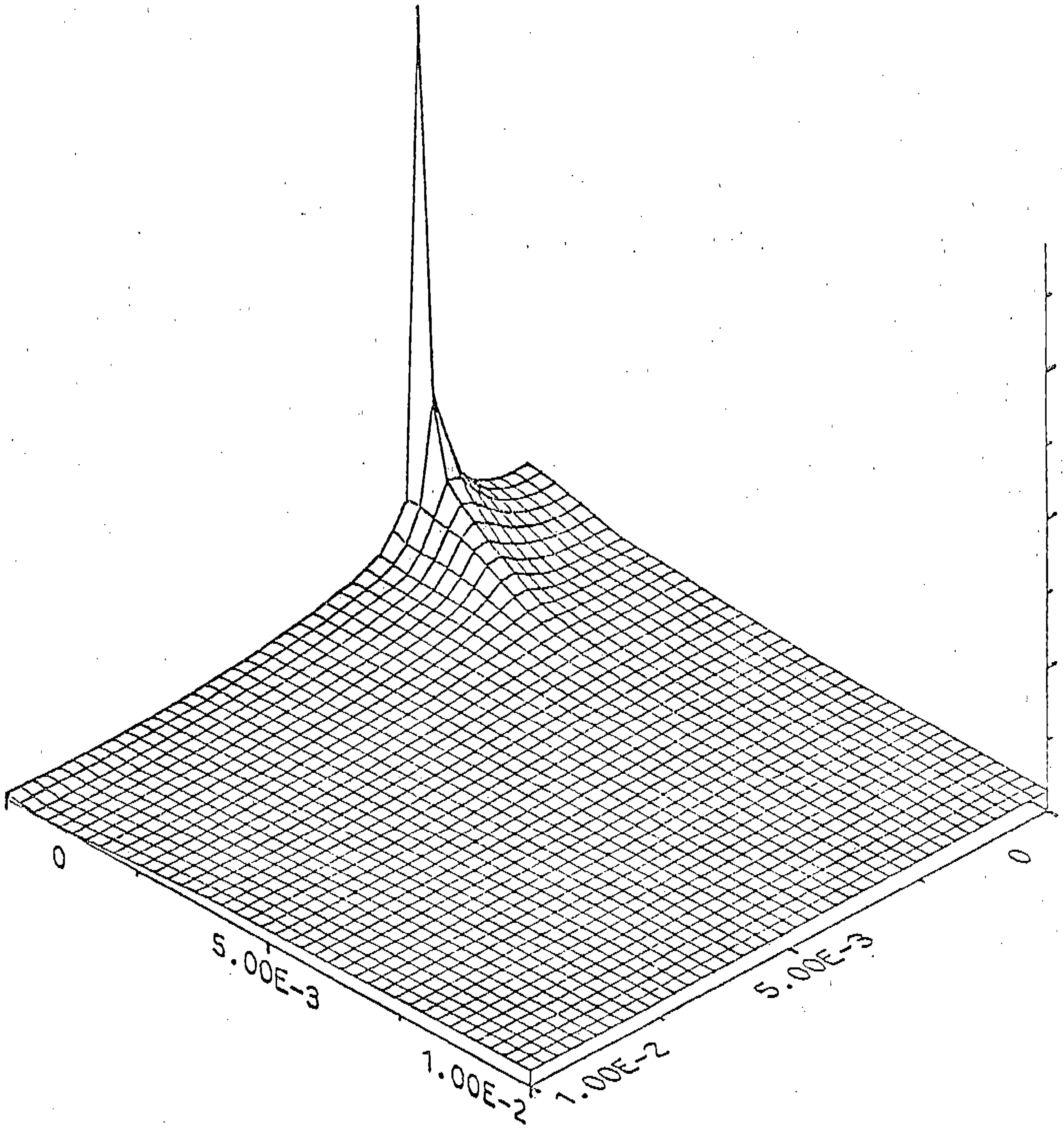


Fig. 10 Electron Current Density (0.7V) [A/cm^2] (lin)

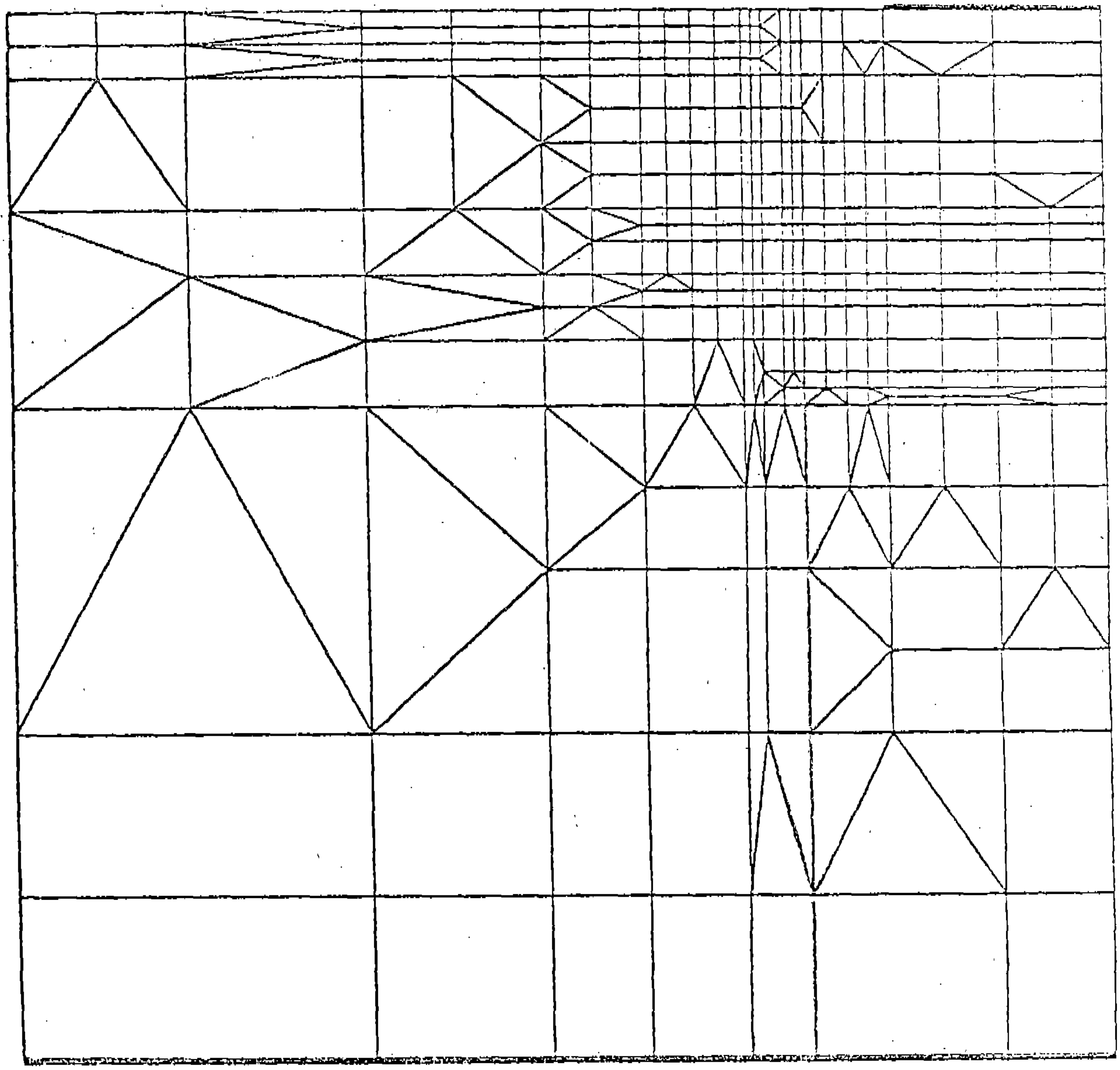


Fig. 11 Final Mesh for 20V Reverse Bias (Finite Elements)

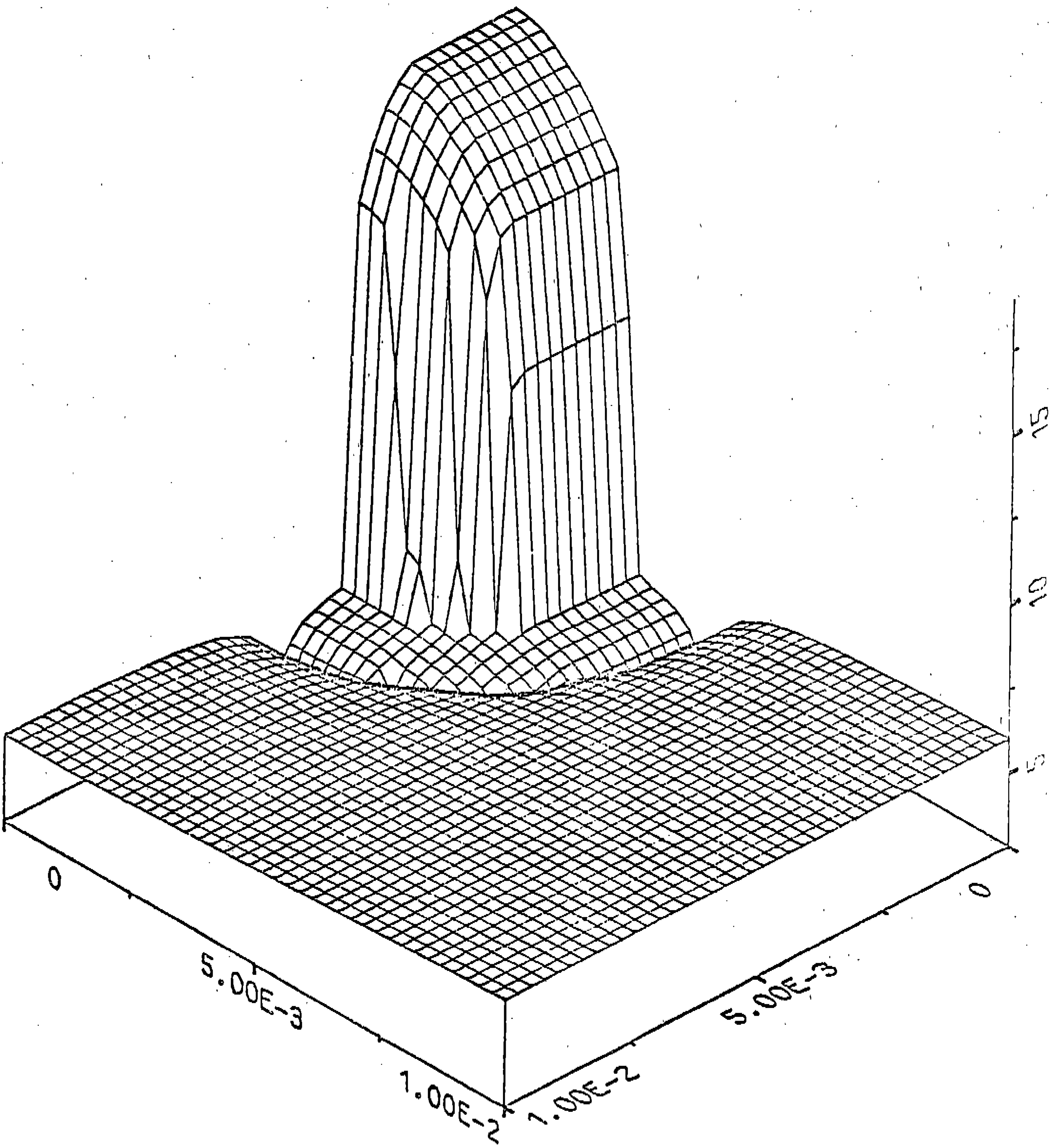


Fig. 12 Electron Concentration (-20V) [cm^{-3}] (log)

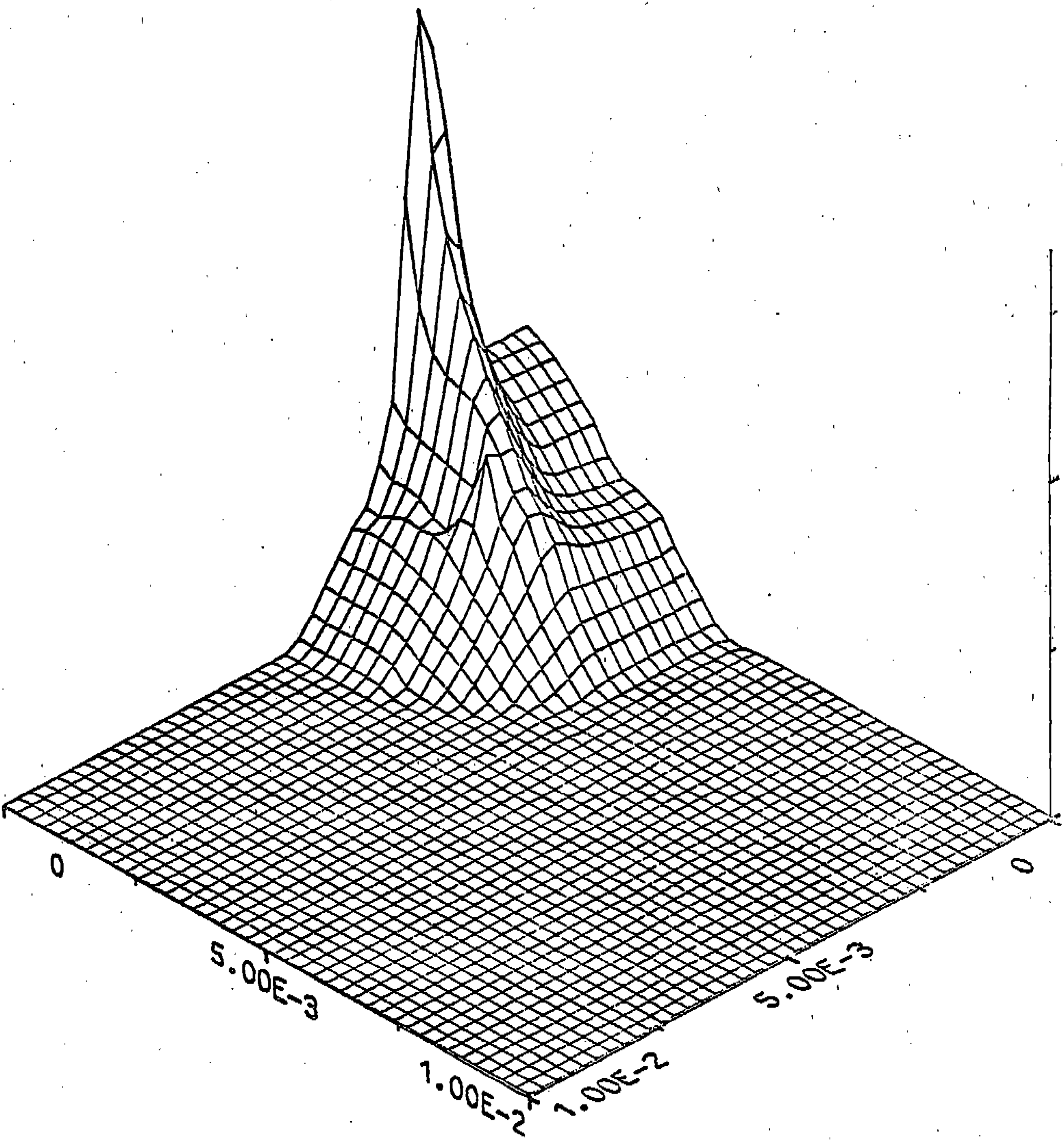


Fig. 13 Electron Current Density (-20V) [A/cm^2] (lin)

References

/1/ Adler M.S.,
"A Method for Achieving and Choosing Variable Density Grids in Finite Difference Formulations and the Importance of Degeneracy and Band Gap Narrowing in Device Modeling",
Proc. NASECODE I Conf., pp.3-30, 1979.

/2/ Adler M.S.,
"A Method for Terminating Mesh Lines in Finite Difference Formulations of the Semiconductor Device Equations",
Solid-State Electron., Vol.23, pp.845-853, 1980.

/3/ Bank R.E., Rose D.J.,
"Parameter Selection for Newton-Like Methods Applicable to Nonlinear Partial Differential Equations",
SIAM J.Numer.Anal., Vol.17, pp.806-822, 1980.

/4/ Bank R.E., Rose D.J.,
"Global Approximate Newton Methods",
Numer. Math., Vol.37, pp.279-295, 1981.

/5/ Bank R.E., Jerome J.W., Rose D.J.,
"Analytical and Numerical Aspects of Semiconductor Device Modeling",
Report 82-11274-2, Bell Laboratories, 1982.

/6/ Birkhoff G.,
"The Numerical Solution of Elliptic Equations",
SIAM, Philadelphia 1971.

/7/ Blatt F.J.,
"Physics of Electronic Conduction in Solids",
McGraw-Hill, New York, 1968.

/8/ Buturla E.M., Cotrell P.E., Grossman B.M., Salsburg K.A., Lawlor M.B.,McMullen C.T,
"Three-Dimensional Finite Element Simulation of Semiconductor Devices",
Proc. International Solid-State Circuits Conf., pp.76-77, 1980.

/9/ Buturla E.M., Cotrell P.E.,
"Simulation of Semiconductor Transport Using Coupled and Decoupled Solution Techniques",
Solid-State Electron., Vol.23, pp.331-334, 1980.

/10/ Buturla E.M., Cottrell P.E., Grossman B.M., Salsburg K.A.,
"Finite-Element Analysis of Semiconductor Devices: The FIELDAY Program",
IBM J. Res. Dev., Vol.25, pp.218-231, 1981.

/11/ Chamberlain S.G., Husain A.,
"Three-Dimensional Simulation of VLSI MOSFET's",
Proc. International Electron Devices Meeting, pp.592-595, 1981.

/12/ Cotrell P.E., Buturla E.M.,
"Two-Dimensional Static and Transient Simulation of Mobile Carrier Transport in a Semiconductor",
Proc. NASECODE I Conf., pp.31-64, 1979.

/13/ De Mari A.,
"An Accurate Numerical One-Dimensional Solution of the P-N Junction under Arbitrary Transient Conditions",
Solid-State Electron., Vol.11, pp.1021-2053, 1968.

/14/ De Mari A.,
"An Accurate Numerical Steady-State One-Dimensional Solution of the P-N Junction",
Solid-State Electron., Vol.11, pp.33-58, 1968.

/15/ Deuflhard P.,
"A Modified Newton Method for the Solution of Ill-Conditioned Systems of Nonlinear Equations with Application to Multiple Shooting",
Numer. Math., Vol.22, pp.289-315, 1974.

/16/ Deuflhard P., Heindl G.,
"Affine Invariant Convergence Theorems for Newton's Method and Extensions to Related Methods",
SIAM J.Numer.Anal., Vol.16, pp.1-10, 1979.

/17/ Doolan E.P., Miller J.J.H., Schilders W.H.A.,
"Uniform Numerical Methods for Problems with Initial and Boundary Layers",
Boole Press, Dublin, 1980.

/18/ Duff I.S.,
"A Survey of Sparse Matrix Research",
Proc. IEEE, Vol.65, pp.500-535, 1977.

/19/ Duff I.S.,
"Practical Comparison of Codes for the Solution of Sparse Linear Systems",
A.E.R.E. Harwell, Oxfordshire 1979.

/20/ Dupont T., Kendall R.D., Rachford H.H.,
"An Approximate Factorization Procedure for Solving Self-Adjoint Elliptic Difference Equations",
SIAM J. Num. Anal., Vol.5, pp.559-573, 1968.

/21/ Engl W.L., Dirks H.,
"Numerical Device Simulation Guided by Physical Approaches",
Proc. NASECODE I Conf., pp.65-93, 1979.

/22/ Fife P.C.,
"Semilinear Elliptic Boundary Value Problems with Small Parameters",
Arch. Rat. Mech. Anal., Vol.29, pp.1-17, 1973.

/23/ Fox L.,
"Finite-Difference Methods in Elliptic Boundary-Value Problems",
in: The State of the Art in Numerical Analysis, pp.799-881, Academic Press, London 1977.

/24/ Franz A.F., Franz G.A., Selberherr S., Ringhofer Ch., Markowich P.,
"FINITE BOXES - a Generalization of the Finite Difference Method Suitable for Semiconductor Device Simulation",
presented at the IEEE/SIAM conf. Numerical Simulation of VLSI Devices, subm. for publ. to IEEE/SIAM, 1982.

/25/ Frey J.,
"Physics Problems in VLSI Devices",
in: Introduction to the Numerical Analysis of Semiconductor Devices and Integrated Circuits, pp.47-50, Boole Press, Dublin 1981.

/26/ Frey J.,
"Transport Physics for VLSI",
in: Introduction to the Numerical Analysis of Semiconductor Devices and Integrated Circuits, pp.51-57, Boole Press, Dublin 1981.

/27/ Greenfield J.A., Price C.H., Dutton R.W.,
"Analysis of Nonplanar Devices",
NATO ASI on Process and Device Simulation for MOS-VLSI Circuits, 1982.

/28/ Grimes R.G., Kincaid D.R., Young D.R.,
"ITPACK 2A - A Fortran Implementation of Adaptive Accelerated Iterative Methods for Solving Large Sparse Linear Systems",
Vol.CNA-164, University of Texas, Austin 1980.

/29/ Gummel H.K.,
"A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations",
IEEE Trans. Electron Devices, Vol.ED-11, pp.455-465, 1964.

/30/ Hess K.,
"Ballistic Electron Transport in Semiconductors",
IEEE Trans. Electron Devices, Vol.ED-28, pp.937-940, 1981.

/31/ Heywang W., Poetzl H.W.,
"Bandstruktur und Stromtransport",
Springer, Berlin 1976.

/32/ Kahng D., Atalla M.M.,
"Silicon-Silicon Dioxide Field Induced Surface Devices",
Solid-State Device Res. Conf., Vol.IRE-AIEE, 1960.

/33/ Kellog R.B.,
"Analysis of a Difference Approximation for a Singular Perturbation Problem in Two Dimensions",
Proc. BAIL I Conf., pp.113-118, Boole Press, Dublin, 1980.

/34/ Kellog R.B., Han Houde,
"The Finite Element Method for a Singular Perturbation Problem Using Enriched Subspaces",
Report BN-978, University of Maryland, 1981.

/35/ Kennedy D.P., O'Brien R.R.,
"Two-Dimensional Mathematical Analysis of a Planar Type Junction Field-Effect Transistor",
IBM J. Res. Dev., Vol.13, pp.662-674, 1969.

/36/ Machek J., Selberherr S.,
"A Novel Finite-Element Approach to Device Modelling",
presented at the IEEE/SIAM conf. Numerical Simulation of VLSI Devices, subm. for publ. to IEEE/SIAM, 1982.

/37/ Manck O., Engl W.L.,
"Two-Dimensional Computer Simulation for Switching a Bipolar Transistor out of Saturation",
IEEE Trans. Electron Devices, Vol.ED-24, pp.339-347, 1975.

/38/ Marhsak A.H., Shrivastava R.,
"Law of the Junction for Degenerate Material with Position-Dependent Band Gap and Electron Affinity",
Solid-State Electron., Vol.22, pp.567-571, 1979.

/39/ Markowich P.A., Ringhofer Ch.A.,
"An Asymptotic Analysis of Single PN-Junction Devices",
Report xxxx, Mathematics Research Center, University of Wisconsin, 1982.

/40/ Markowich P.A., Ringhofer Ch.A., Selberherr S., Langer E.,
"A Singularly Perturbed Boundary Value Problem Modelling a Semiconductor Device",
Report 2388, Mathematics Research Center, University of Wisconsin, 1982.

/41/ Markowich P.A., Ringhofer Ch.A., Selberherr S.,
"A Singular Perturbation Approach for the Analysis of the Fundamental Semiconductor Equations",
presented at the IEEE/SIAM conf. Numerical Simulation of VLSI Devices, subm. for publ. to IEEE/SIAM, 1982.

/42/ Markowich P.A.,
"Zur zweidimensionalen Analyse der Halbleitergrundgleichungen",
Habilitation, Technical University of Vienna, 1983.

/43/ Mock M.S.,
"On Equations Describing Steady-State Carrier Distributions in a Semiconductor Device",
Comm. Pure and Appl. Math., Vol.25, pp.781-792, 1972.

/44/ Mock M.S.,
"A Time-Dependent Numerical Model of the Insulated-Gate Field-Effect Transistor",
Solid-State Electron., Vol.24, pp.959-966, 1981.

/45/ Moglestue C., Beard S.J.,
"A Particle Model Simulation of Field Effect Transistors",
Proc. NASECODE I Conf., pp.232-236, 1979.

/46/ Moglestue C.,
"A Monte-Carlo Particle Model Study of the Influence of the Doping Profiles on the Characteristics of Field-Effect Transistors",
Proc. NASESCODE II Conf., pp.244-249, 1981.

/47/ Nussbaum A.,
"Inconsistencies in the Original Form of the Fletcher Boundary Conditions",
Solid-State Electron., Vol.21, pp.1178-1179, 1978.

/48/ Ortega J.M., Rheinboldt W.C.,
"Iterative Solution of Nonlinear Equations in Several Variables",
Academic Press, New York 1970.

/49/ Price C.H.,
"Two-Dimensional Numerical Simulation of Semiconductor Devices",
Dissertation, Stanford University, 1980.

/50/ Scharfetter D.L., Gummel H.K.,
"Large-Signal Analysis of a Silicon Read Diode Oscillator",
IEEE Trans. Electron Devices, Vol.ED-16, pp.64-77, 1969.

/51/ Schuetz A., Selberherr S., Poetzl H.W.,
"A Two-Dimensional Model of the Avalanche Effect in MOS Transistors",
Solid-State Electron., Vol.25, pp.177-183, 1982.

/52/ Seeger K.,
"Semiconductor Physics",
Springer, Wien 1973.

/53/ Selberherr S., Schuetz A., Poetzl H.W., "MINIMOS - a Two-Dimensional MOS Transistor Analyzer", IEEE Trans. Electron Devices, Vol.ED-27, pp.1540-1550, 1980.

/54/ Selberherr S., Schuetz A., Poetzl H., "Two Dimensional MOS-Transistor Modeling", NATO ASI on Process and Device Simulation for MOS-VLSI Circuits, 1982.

/55/ Smith G.D., "Numerical Solution of Partial Differential Equations: Finite Difference Methods", Clarendon Press, Oxford 1978.

/56/ Smith R.A., "Semiconductors", Cambridge University Press, Cambridge 1978.

/57/ Stoer J., Bulirsch R., "Einfuehrung in die Numerische Mathematik II", Springer, Berlin 1978.

/58/ Stone H.L., "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations", SIAM J.Numer.Anal., Vol.5, pp.530-558, 1968.

/59/ Toyabe T., Yamaguchi K., Asai S., Mock M., "A Numerical Model of Avalanche Breakdown in MOSFET's", IEEE Trans. Electron Devices, Vol.ED-25, pp.825-832, 1978.

/60/ Van Overstraeten R.J., De Man H.J., Mertens R.P., "Transport Equations in Heavy Doped Silicon", IEEE Trans. Electron Devices, Vol.ED-20, pp.290-298, 1973.

/61/ Van Roosbroeck W.V., "Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors", Bell Syst. Techn. J., Vol.29, pp.560-607, 1950.

/62/ Van Vliet K.M., "On Fletcher's Boundary Conditions", Solid-State Electron., Vol.22, pp.443-444, 1979.

/63/ Van Vliet K.M., "The Shockley-Like Equations for the Carrier Densities and the Current Flows in Materials with a Nonuniform Composition", Solid-State Electron., Vol.23, pp.49-53, 1980.

/64/ Varga R.S.,
"Matrix Iterative Analysis",
Prentice-Hall, Englewood Cliffs 1962.

/65/ Vasilev'a A.B., Butuzov V.F.,
"Singularly Perturbed Equations in the Critical Case",
translated Report 2039, Mathematics Research Center,
University of Wisconsin, 1978.

/66/ Wachspress E.L.,
"Iterative Solution of Elliptic Systems",
Prentice-Hall, Englewood Cliffs 1966.

/67/ Yoshii A., Horiguchi S., Sudo T.,
"A Numerical Analysis for Very Small Semiconductor Devices",
Proc. International Solid-State Circuits Conf., pp.80-81, 1980.

/68/ Yoshii A., Kitazawa H., Tomizawa M., Horiguchi S., Sudo T.,
"A Three-Dimensional Analysis of Semiconductor Devices",
IEEE Trans. Electron Devices, Vol.ED-29, pp.184-189, 1982.

CR/SS/jvs

NON-LINEAR MULTIDIMENSIONAL MODELS, LEAST SQUARES AND REDUCED RESIDUALS

Aivars Celmiņš
Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. This paper addresses the treatment of such model fitting problems that include an analysis of the residuals, for instance to determine heteroscedasticity. In cases with multidimensional observations the residuals are vectors having model induced correlations between components. It is shown that one can eliminate such correlations and also reduce the number of residual components by introducing a concept of reduced residuals. The application of the new concept is illustrated by an investigation of heteroscedasticity of vapor pressure measurements.

1. INTRODUCTION. The residuals of model fitting problems are valuable sources of information about observational errors, model adequateness and the relative importance of observations. The information is typically obtained by analyzing the distribution and other properties of the residuals. If the model equation is formulated in terms of one scalar dependent variable, and the independent variables are assumed to be error free then the analysis involves only standard procedures for the distribution of a scalar quantity. If, however, the adjustable observations are not scalars, for instance, if the independent variables are subject to errors, then also the residuals are vectors and the usual methods of residual investigation must be modified. This paper addresses such modifications in cases of general non-linear models, which are assumed to be formulated by systems of implicit equations, and fitted by least squares. We show that the n-dimensional residuals of the observations are not necessarily appropriate for an analysis of distribution, heteroscedasticity and other properties, because the constraints induce correlations between residual components. We, therefore, introduce the concept of reduced residuals. These residuals have a dimension that is less or equal to that of the observables, and their components are dimensionless and free from model induced correlations. In standard least squares problems, where the model equations are scalar, also the reduced residuals are scalars and standard techniques can be used for the investigation of their distribution. One also can show that no information is lost by analysing the reduced residuals instead of the original residuals, that is, the reduced residuals contain all the information about the real errors that can be extracted from the model fitting. An example of the use of reduced residuals is presented by discussing the investigation of heteroscedasticity of vapor pressure measurements.

The main definitions and principal results are presented in Section 2, and an example of applications is given in Section 3. The background of the theory is outlined in the Appendix, where we present detailed derivations of the formulas of Section 2.

2. REDUCED RESIDUALS IN GENERAL MODEL FITTING PROBLEMS.

We give in this section the definition of reduced residuals and present some of their more important properties. Derivations of the formulas and detailed discussions are provided in the Appendix.

We consider least squares model fitting problems that can be formulated as the following constrained minimization task:

$$\text{minimize} \qquad W = \sum_{i=1}^{s} c_i^T R_i^{-1} c_i, \tag{2.1a}$$

$$\text{subject to} \qquad F_i(X_i + c_i, t) = 0,\ i = 1,\ldots,s \quad . \tag{2.1b}$$

The X_i are s observed points in a n-dimensional space of observables, that is, each observation X_i is a n-dimensional vector; the c_i are the corresponding residual vectors; t is a p-dimensional vector of model parameters; and R_i are estimated (n x n)-dimensional variance-covariance matrices of the observations X_i. Each constraint equation (2.1b) is a set of r equations, that is, the model functions $F_i(\xi,t)$ are r-dimensional vector functions of n + p variables. The problem (2.1) includes as a special case elementary weighted least squares problems for which r = 1, n = 1 and the functions $F_i(\xi,t)$ are linear with respect to ξ. The unknowns of the problem (2.1) are the residual vectors c_i and the model parameter vector t. The given input consists of the observations X_i, their estimated variances and covariances R_i, and postulated functional relationships $F_i(\xi,t) = 0$.

The least squares residuals c_i are n-dimensional vectors but they do not necessarily span a n-dimensional space. For instance, if one fits a straight line to observations in a plane, whereby both coordinates are adjusted and the R_i are all equal, then all residual vectors c_i are parallel to each other. That is, they are elements of a one-dimensional subspace (line) of the two-dimensional space of observables. In this example all information about distribution properties of the residuals is contained in the algebraic lengths of the residuals, and the analysis of the residuals can be reduced to the analysis of the distribution of a scalar.

In order to effectively handle the described situation and similar more general problems we define reduced residuals a_i by the equation

$$a_i = (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi} c_i \quad , \tag{2.2}$$

where F_{xi} is the Jacobian matrix

$$F_{xi} = \frac{\partial F_i(X_i + c_i,\ t)}{\partial X_i} \quad . \tag{2.3}$$

The reduced residuals a_i have the same number of components as F_i, and the components are dimensionless. In the planar curve fitting problem described above, the reduced residuals a_i are dimensionless scalars with an absolute value equal to a norm of c_i, and positive or negative, depending on which side of the line $F_i = 0$ the observation X_i is located.

Proper norms of the residuals c_i are in least squares problems the elliptic norms

$$||c_i|| = (c_i^T R_i^{-1} c_i)^{1/2}, \tag{2.4}$$

because the objective function W in (2.1a) then is the sum of the squares of the norms $||c_i||$. For later reference we also define an inner product associated with each observation X_i by

$$(c_i, b_i) = c_i^T R_i^{-1} b_i \; . \tag{2.5}$$

We show in the Appendix that the Euclidean norm of the r-dimensional reduced residual a_i is equal to the elliptic norm (2.4) of the n-dimensional residual c_i:

$$||a_i|| = (a_i^T a_i)^{1/2} = (c_i^T R_i^{-1} c_i)^{1/2} = ||c_i||. \tag{2.6}$$

In a standard least squares problem the F_i are scalar model functions. In that case the definition (2.2) of the reduced residuals can be replaced by

$$a_i = (c_i^T R_i^{-1} c_i)^{1/2} \text{ sgn } (F_{xi} \; c_i). \tag{2.7}$$

In elementary least squares problems the F_i are assumed to be linear with respect to the observable. In that case Eq. (2.7) futher simplifies to

$$a_i = c_i/e_i, \tag{2.8}$$

where e_i is the estimated standard error of the scalar observation X_i. Hence in elementary least squares problems the reduced residuals are identical to the familiar weighted residuals.

The concept of reduced residuals can be geometrically interpreted as follows. The fitted r-dimensional model equation $F_i(\xi,t) = 0$ defines in the n-dimensional ξ-space of observables a (n - r)-dimensional hypersurface, and the corrected observation $X_i + c_i$ is a point of that surface. The least squares residual c_i is a n-dimensional vector orthogonal to the hypersurface $F_i = 0$. (Orthogonal in the sense of the inner product (2.5).) That is, all possible least squares residuals corresponding to the hypersurface point $X_i + c_i$ define a hyperplane orthogonal to the surface $F_i = 0$. The hyperplane is a r-dimensional linear subspace of the n-dimensional ξ-space, and the reduced residuals a_i are elements of that subspace.

The advantages of working with the a_i instead of the c_i are as follows. First, the a_i have in general less components than the c_i. Second, the components of the a_i are dimensionless whereas the components of the c_i have generally different physical dimensions. Third, there are no restrictions on the components of the a_i, whereas the components of the c_i are restricted by the condition of orthogonality to $F_i = 0$. The latter condition induces apparent correlations between components of the c_i; the components of the a_i are free from such correlations. If correlations between components of the a_i are detected then they generally indicate correlations between observations as we will show next.

A first order relation between the reduced residual a_i and the unknown real error $\overset{*}{c}_i$ of the observation X_i is (see Appendix, Eq. (A.23))

$$a_i = (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi}\overset{*}{c}_i + (F_{xi} R_i F_{xi}^T)^{-1/2} F_{ti} (\overset{*}{t} - t), \qquad (2.9)$$

where F_{ti} is the Jacobian matrix

$$F_{ti} = \frac{\partial F_i(X_i + c_i,\ t)}{\partial t} \qquad (2.10)$$

and $\overset{*}{t}$ is the unknown true value of the model parameter. The relation (2.9) neglects terms of higher order in $\overset{*}{c}_i - c_i$ and $\overset{*}{t} - t$. The second term on the right hand side of Eq. (2.9) is a function that in general slowly varies along the hypersurface $F_i = 0$. It can be interpreted as a distance between the true surface $F_i(\xi,\overset{*}{t}) = 0$ and the least squares fit $F_i(\xi,t) = 0$. The first term on the right hand side is the projection of the true observational error $\overset{*}{c}_i$ on the subspace orthogonal to $F_i = 0$. Therefore, any local scatter of the true observational errors will usually manifest itself as a local scatter of the reduced residuals a_i. In particular, if the $\overset{*}{c}_i$ are componentwise normally distributed then so are their projections and, except for a bias caused by the slowly varying second term in Eq. (2.9), the reduced residuals a_i.

Figure 1 presents a geometrical interpretation for the case where $F_i = 0$ defines a curve in the plane of observations. The least squares residual c_i is orthogonal in the sense of the inner product (2.5) to the fitted curve, and it defines the subspace (line) for a_i. The latter is a measure for the algebraic length of c_i along the normal. According to Eq. (2.9) a_i is the sum of the projection of the true error c_i^* on the normal and a term representing the distance between the true and the fitted curve, measured along the same normal.

3. APPLICATION EXAMPLE. An example for the use of reduced residuals is the following investigation of heteroscedasticity of vapor pressure measurements. The measurements consist of a series of pressure-temperature correspondences to which one fits the so-called Antoine equation, viz.,

$$F(p,T;\ A,B,C) = \lg(p/p_R) - A + B/[(T - 273.15) + C] = 0 \qquad (3.1)$$

where p (Pa) is pressure, T (K) is temperature, p_R (Pa) is a reference pressure (usually p_R = 1 torr = $7.50064 \cdot 10^{-5}$ Pa), and A, B and C are model parameters [1]. Because the model function (3.1) is scalar, this is a standard least squares problem that can be solved numerically with available utility routines [2] provided that estimates are available of the accuracies of the observations of T and p. More typical for these measurements is a situation where the information about data accuracies is incomplete: one can assume that the temperature measurements are all made with the same standard error, but the pressure standard errors likely do depend on the pressure. Then the problem is to find from the same data set estimates for the model parameters A, B and C as well as estimates for other parameters describing the data accuracy, particularly the dependence of the pressure accuracy on the pressure. We assume that all observations are independent and postulate the following models for the standard errors of pressure and temperature measurements:

$$e_{pi}/p_R = e_o\ (1 + \theta_1(p_i + c_{pi})/p_R)$$
$$e_{Ti} = e_o\theta_2 \qquad (3.2)$$

In Eq. (3.2) e_o is the standard error of weight one which is computed after the adjustment by

$$e_o = [W/(s - 3)]^{1/2}, \qquad (3.3)$$

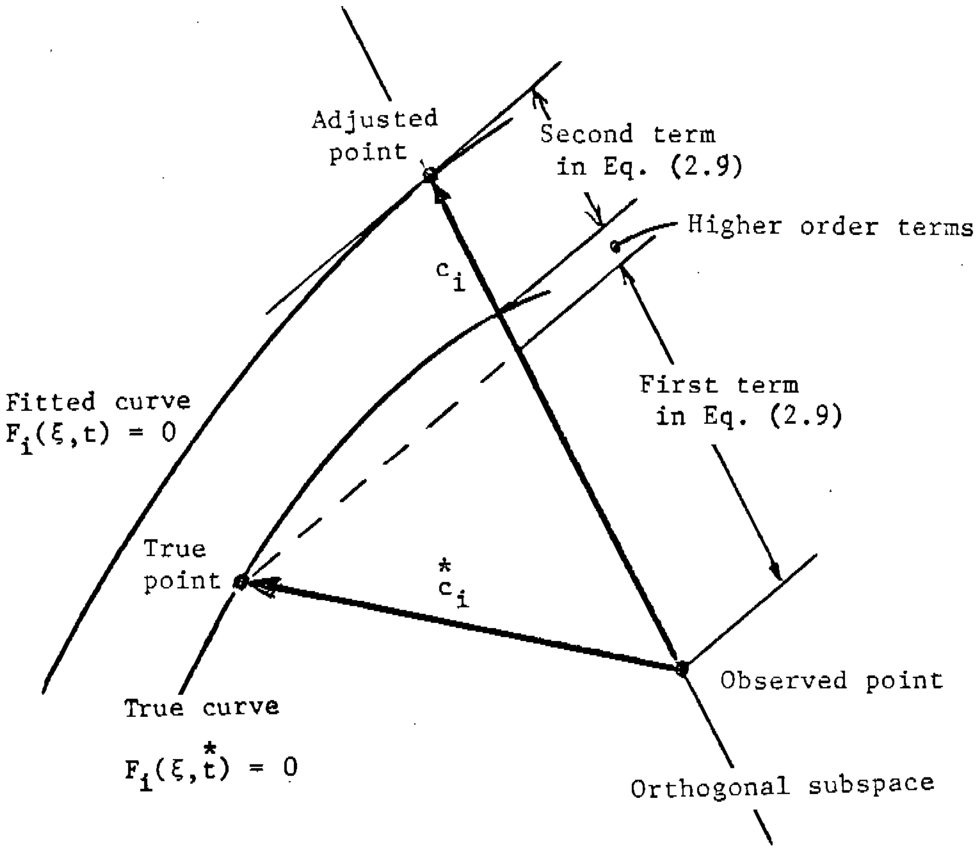


Figure 1. Reduced Residual in Planar Curve Fitting.

The reduced residual a_i is a scalar measuring the algebraic length of the residual c_i.

and c_{pi} is the least squares residual of the observed pressure p_i. The error model (3.2) contains two parameters, θ_1 and θ_2. For small θ_1 it represents a constant pressure error assumption, and for large θ_1 it represents a constant relative pressure error. The second error model parameter θ_2 has the dimension of temperature and it permits one to model the relative significance of pressure and temperature observations. If $\theta_2 = 0$ then the temperature observations are assumed to be error free.

We determine the error model parameters by minimizing an objective function S which we define in terms of the reduced residuals as follows

$$S = 1 + \frac{1}{\ln s} \sum_{i=1}^{s} q_i \ln q_i , \qquad (3.4a)$$

where

$$q_i = ||a_i||^2 / \sum_{i=1}^{s} ||a_i||^2 \quad . \qquad (3.4b)$$

The use of the negative entropy function S as an objective function was suggested by Nielsen [3]. His definition of the q_i in Eq. (3.4b) was, however, in terms of weighted residuals instead of reduced residuals, because he was only treating elementary weighted least squares problems. Nielsen chose S as a measure for the optimality of the error model parameters because of the following properties of the function:

(a) S is maximum and equals one if all but one of the q_i are zero.

(b) S is minimum and equals zero of all $q_i = 1/s$.

(c) Any averaging of the q_i reduces S, that is, if

$$0 \leqslant a_{ij} \leqslant 1, \qquad \sum_{i=1}^{s} a_{ij} = \sum_{j=1}^{s} a_{ij} = 1, \qquad i = 1,2,\ldots,s$$

and

$$\overset{*}{q}_i = \sum_{i=1}^{s} a_{ij}\, q_j ,$$

then

$$S\,(\overset{*}{q}) \leqslant S\,(q) \quad .$$

Hence a smaller S generally means smaller differences between the q_i. Because of the definition (3.4b) and the relation (2.6) this in turn means that a minimization of S tends to equalize the $||c_i||^2$. For given values of θ_1 and θ_2 one computes S by first carrying out a least squares adjustment with the model function (3.1) and the error estimates (3.2). The adjustment produces least squares values of model parameters A, B and C and a set of residual vectors c_i. These results provide via Eq. (2.7) the reduced residuals from which the value of S can be computed by Eq. (3.4).

We notice in passing that the choice of S as an objective function is arbitrary and one may instead use other functions with similar properties [4]. In limited numerical experiments we indeed found little difference between corresponding results with different objective functions. The results quoted in this paper are for the objective function S as defined by Eq. (3.4).

The optimality of the error model parameters can also be defined and tested by other means than an objective function. For instance, one can in the present example investigate the distribution of the reduced residuals and compare it to a normal distribution if the real errors are known to be normally distributed with a heteroscedasticity parameter θ_1 in the form of Eq. (3.2). Then the optimal values of θ_1 and θ_2 would be those for which the reduced residual distribution is closest to normal according to some appropriate criterium. As we shall see in our examples, such an approach is practically equivalent to the minimization of S.

Explicit formulas for the reduced residuals and other quantities in the case of vapor pressure measurements are as follows. The observation vectors X_i are

$$X_i = \begin{pmatrix} p_i \\ T_i \end{pmatrix} \tag{3.5}$$

the corresponding residuals are

$$c_i = \begin{pmatrix} c_{pi} \\ c_{Ti} \end{pmatrix}, \tag{3.6}$$

the estimated variance-covariance matrices are

$$R_i = \begin{pmatrix} p_R^2 \, [1 + \theta_1(p_i + c_{pi})/p_R]^2 & 0 \\ 0 & \theta_2^2 \end{pmatrix} \tag{3.7}$$

and the reduced residuals are

$$a_i = [(\frac{c_{pi}/p_R}{1 + \theta_1(p_i + c_{pi})/p_R})^2 + (\frac{c_{Ti}}{\theta_2})^2]^{1/2} \text{ sgn } (F_{pi}c_{pi} + F_{Ti}c_{Ti}) \quad (3.8)$$

where F_{pi} and F_{Ti} are the partial derivatives of the model function (3.1) with respect to p and T, evaluated at the adjusted observations. The constraint equations are obtained from Eq. (3.1) by setting

$$F(p_i + c_{pi}, T_i + c_{Ti}; A, B, C) = 0 \quad , \quad i = 1,\ldots,s. \quad (3.9)$$

Next we present a numerical example with simulated data and known error distribution. The data were obtained by chosing a set A, B, C of Antoine parameters, calculating for s = 40 equidistant $\overset{*}{T}_i$ values the corresponding $\overset{*}{p}_i$ from Eq. (3.1), and subtracting from the $\overset{*}{p}_i$ and $\overset{*}{T}_i$ random errors $\overset{*}{c}_{pi}$ and c_{Ti} with known normal distributions. The simulated observations thus have the values

$$p_i = \overset{*}{p}_i - \overset{*}{c}_{pi} \quad ,$$
$$T_i = \overset{*}{T}_i - \overset{*}{c}_{Ti} \quad . \quad (3.10)$$

Figure 2 shows a typical set of simulated data. The Antoine parameters that were used to calculate $\overset{*}{p}_i(\overset{*}{T}_i)$ and the error model parameters for the calculation of $\overset{*}{c}_{pi}$ and $\overset{*}{c}_{Ti}$ are listed in Table 1. The table also contains the parameters of the fitted curve shown in Figure 2, and of the error ellipses. The curve and the error ellipses correspond to an optimal value of θ_1 that minimizes the objective function S for this data set. The second error model parameter θ_2 was chosen such that $e_T = e_0\theta_2 = 0.1$ K, that is, the temperature standard error e_T was preset to the exact value. The least squares fitting was done using the utility program COLSAC [2]. The confidence limits for the fitting curve were computed by solving the Antoine Eq. (3.1) for p and applying the linearized law of variance propagation to the function p(T;A,B,C), that is, by

$$e_p = [\frac{\partial p(T;A,B,C)}{\partial\,(A,B,C)} V_{ABC} (\frac{\partial p(T;A,B,C)}{\partial(A,B,C)})^T]^{1/2} \quad , \quad (3.11)$$

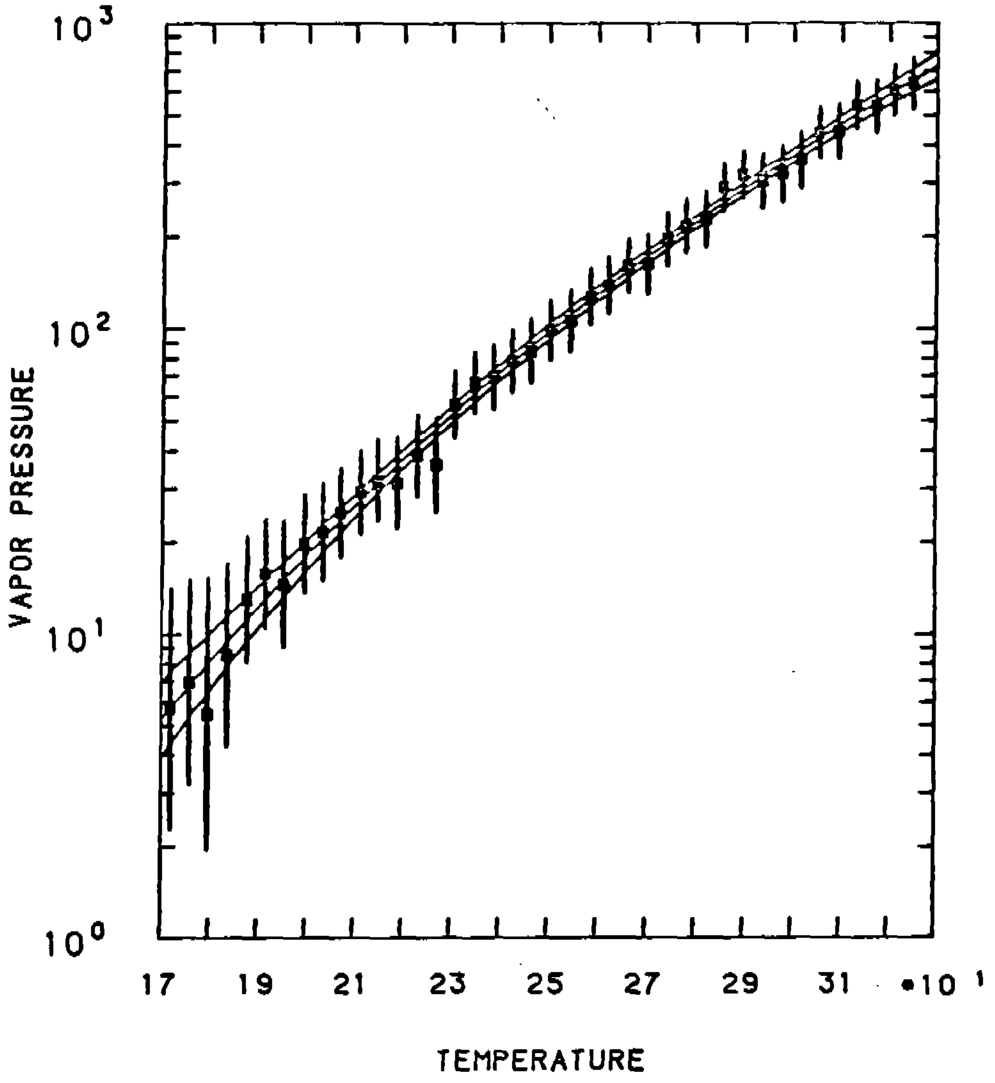


Figure 2. Simulated Data with Fitted Curve.

Error assumptions for the fitting are

$$e_{pi}/p_R = 1.19 + 0.055\ (p_i + c_{pi})/p_R,$$

$$e_{Ti} \quad = 0.1\ \mathrm{K}$$

The confidence limits and error ellipses correspond to 3.4 standard errors. Pressure is shown in torr and temperature is shown °C.

where $\partial p/\partial(A,B,C)$ is the Jacobian matrix of the function $p(T;A,B,C)$ (the gradient of p in the A,B,C-space), and V_{ABC} is the estimated variance-covariance matrix of the parameters A, B and C. The matrix V_{ABC} is defined in terms of the standard errors of A, B and C, and of the corresponding correlation matrix C_{ABC} (all given in Table 1) by

$$V_{ABC} = D_{ABC}\, C_{ABC}\, D_{ABC} \quad , \tag{3.12}$$

where D_{ABC} is a diagonal matrix with the standard errors in the diagonal. About the computation of V_{ABC} (and C_{ABC}) from the least squares adjustment see reference [5]. The matrices are part of the output of the utility program COLSAC.

The dependence of the objective function S on the pressure error parameter θ_1 is illustrated in Figure 3 for a fixed temperature standard error $e_T = 0.1$ K. The shape of the curve is typical for sample problems with a sufficiently large number of data ($s \geqslant 40$) and a reasonable $e_T \leqslant 1.0$ K. Also the numerical results are only little influenced by the preset value of e_T, as shown in Table 1. We conclude from these experiments that the pressure error heteroscedasticity (the parameters e_o and $e_o\theta_1$) can be reasonably well retrieved if $s \geqslant 40$, but that the temperature standard error e_T practically cannot be retrieved. If the number of data points is too small ($s \leqslant 10$) then one observes large variations of the optimal θ_1-values between different sets of random input errors. If the temperature standard error is preset to a high value ($e_T > 1.0$ K) then the least squares adjustment produces a solution were practically only the temperature observations are adjusted and, therefore, the pressure error parameters cannot be retrieved. The difference between input pressure error parameters and retrieved parameters shown in Table 1 is typical for the given size of the problem. The differences and case to case variations decrease if the number of observations increases.

The distribution of the reduced residuals is illustrated by Figures 4 through 6. The figures show the normal distribution compared to the cumulative residual distribution in a case described in Table 1 and for different values of θ_1. It is obvious that smaller values of the objective function S correspond to reduced residual distributions that are closer to normal. For comparison we show in Figure 7 the cumulative distribution of the reduced true input errors for the same example. One observes that the optimal distribution of the reduced least squares residuals (Figure 5) is closer to normal than the distribution of the corresponding reduced true input errors, drawn from a random number generator.

The next example is a case with real data taken from reference [1]. The data are shown in Figure 8 and numerical results are given in Table 2. The dependence of the objective function S on the pressure error parameter θ_1 is illustrated by Figures 9 and 10. In this case $S(\theta)$ has no minimum for a finite θ_1. Therefore, the optimal choice is a constant relative pressure

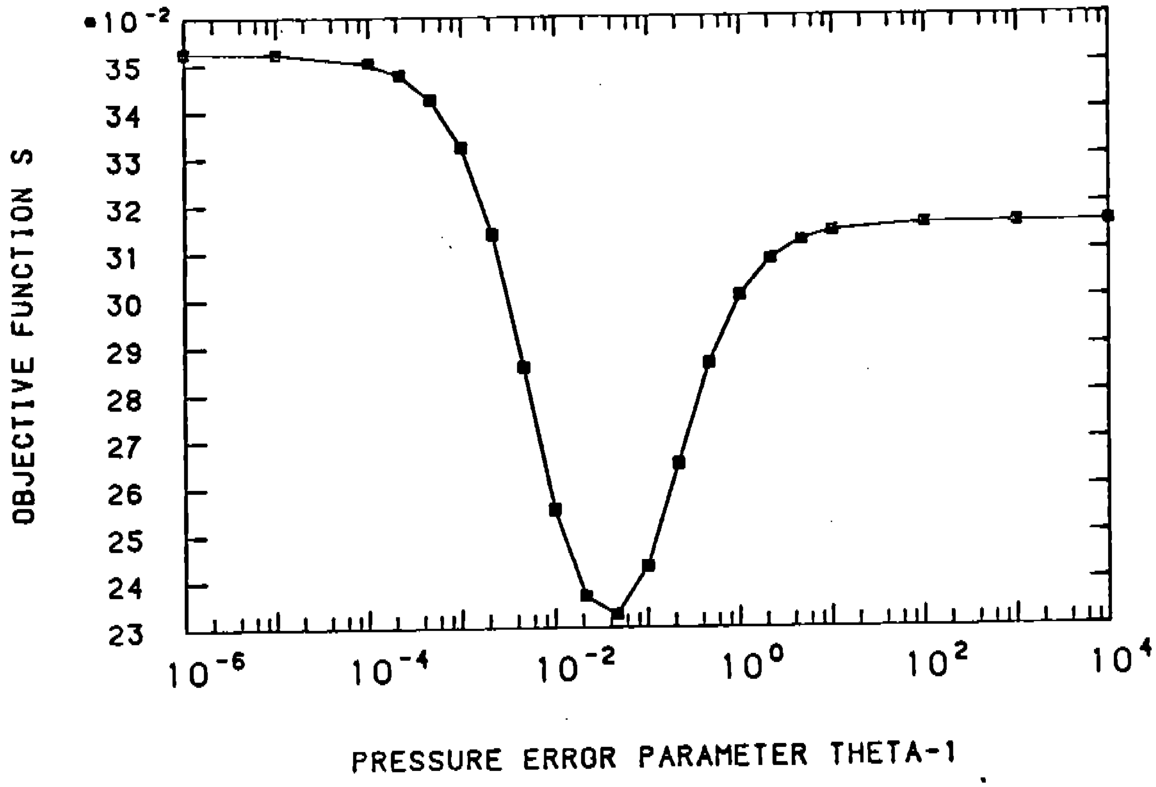


Figure 3. Objective Function S for Simulated Data.

The temperature standard error $e_T = e_o \theta_2$ is set to 0.1 K.

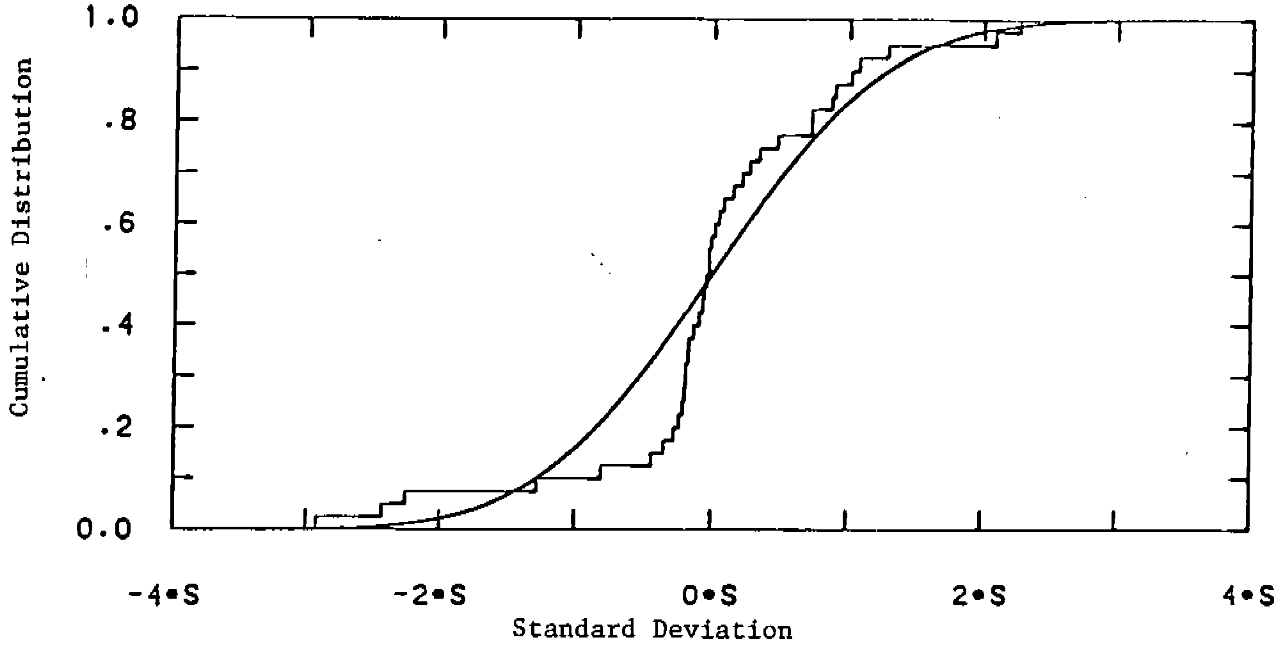


Figure 4. Comulative Distribution of Reduced Residuals for Simulated Data and Small $\theta_1 = 10^{-6}$.

The temperature standard error is set to $e_T = 0.1$ K.

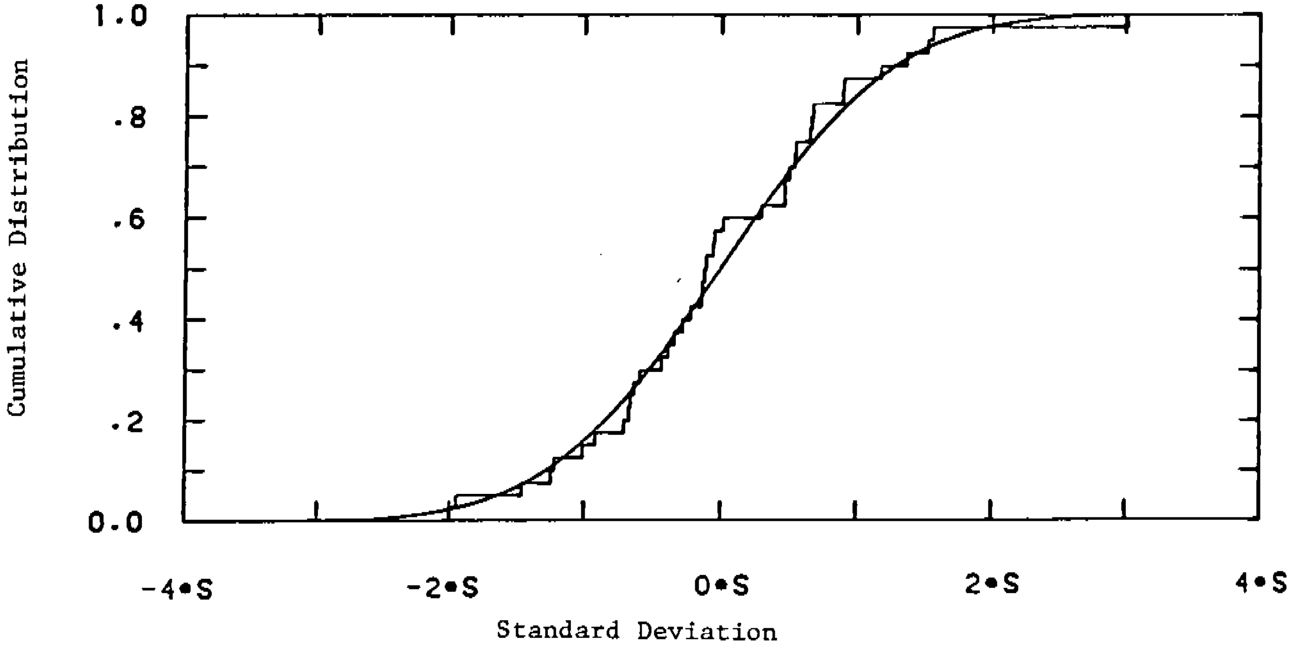


Figure 5. Cumulative Distribution of Reduced Residuals for Simulated Data and Optimal $\theta_1 = 0.049$.

The temperature standard error is set to $e_T = 0.1$ K.

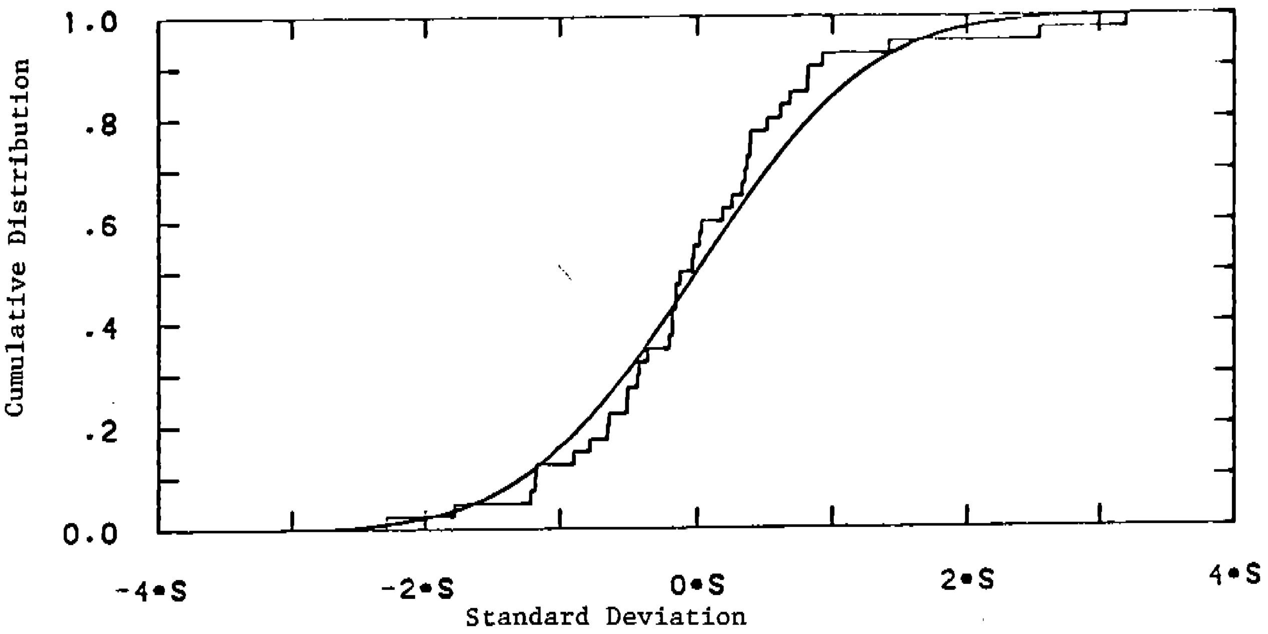


Figure 6. Cumulative Distribution of Reduced Residuals for Simulated Data and Large $\theta_1 = 10^4$.

The temperature standard error is set to $e_T = 0.1$ K.

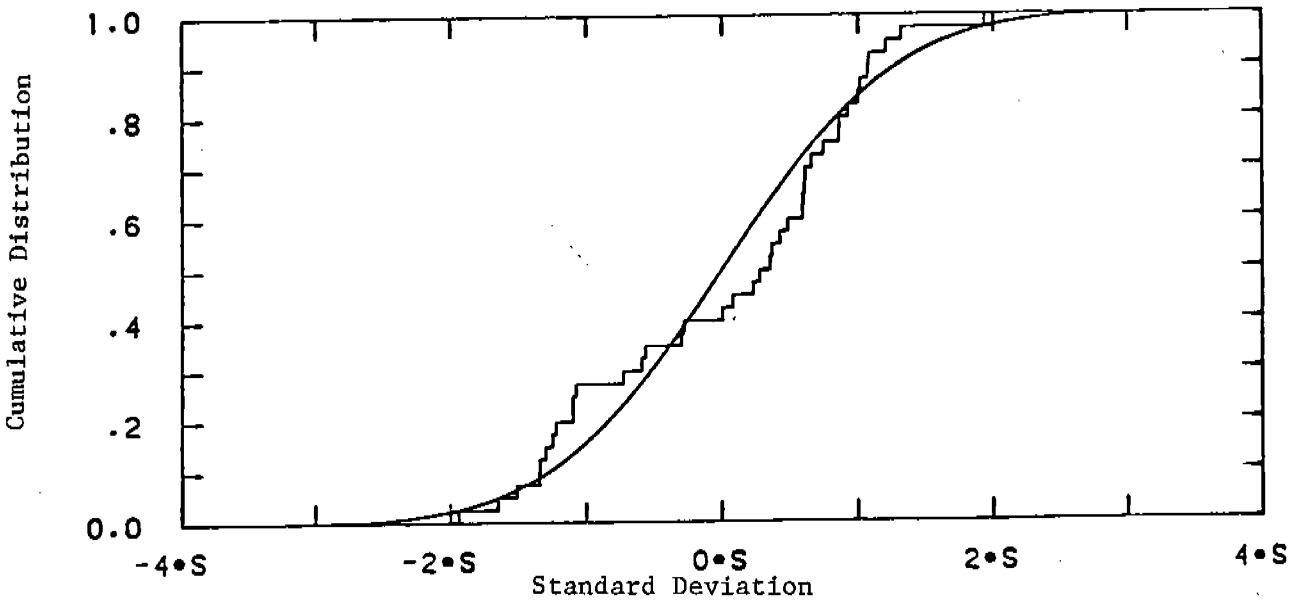


Figure 7. Cumulative Distribution of Reduced True Simulated Errors.

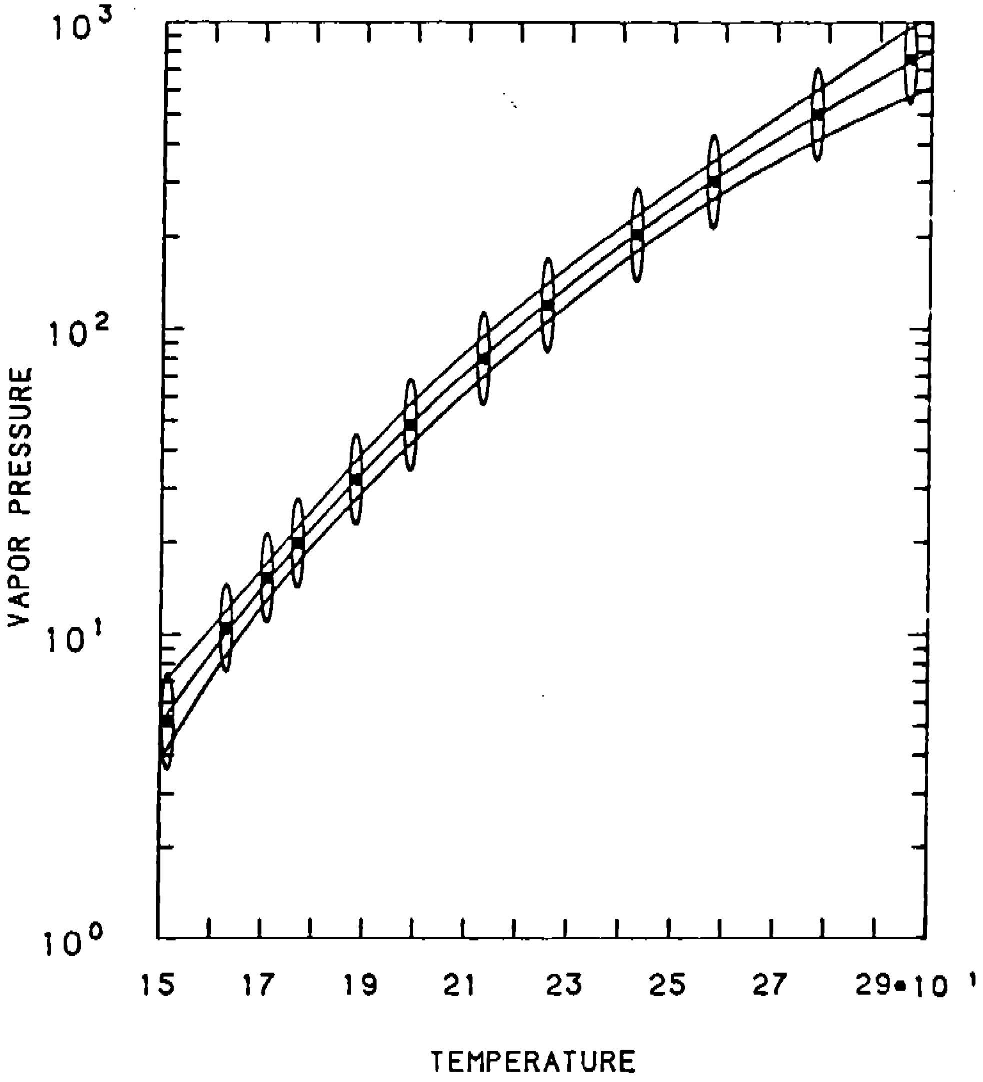


Figure 8. Data and Fitted Curve for 1-Tetradecanol.

Error assumptions Curve for fitting are

$$e_{pi}/p_R = 0.0294\ (p_i + c_{pi})/p_R$$

$$e_{Ti} \quad = 0.1\ K.$$

The confidence limits and error ellipses correspond to 11.9 standard errors. Pressure is shown in torr and temperature is shown in °C.

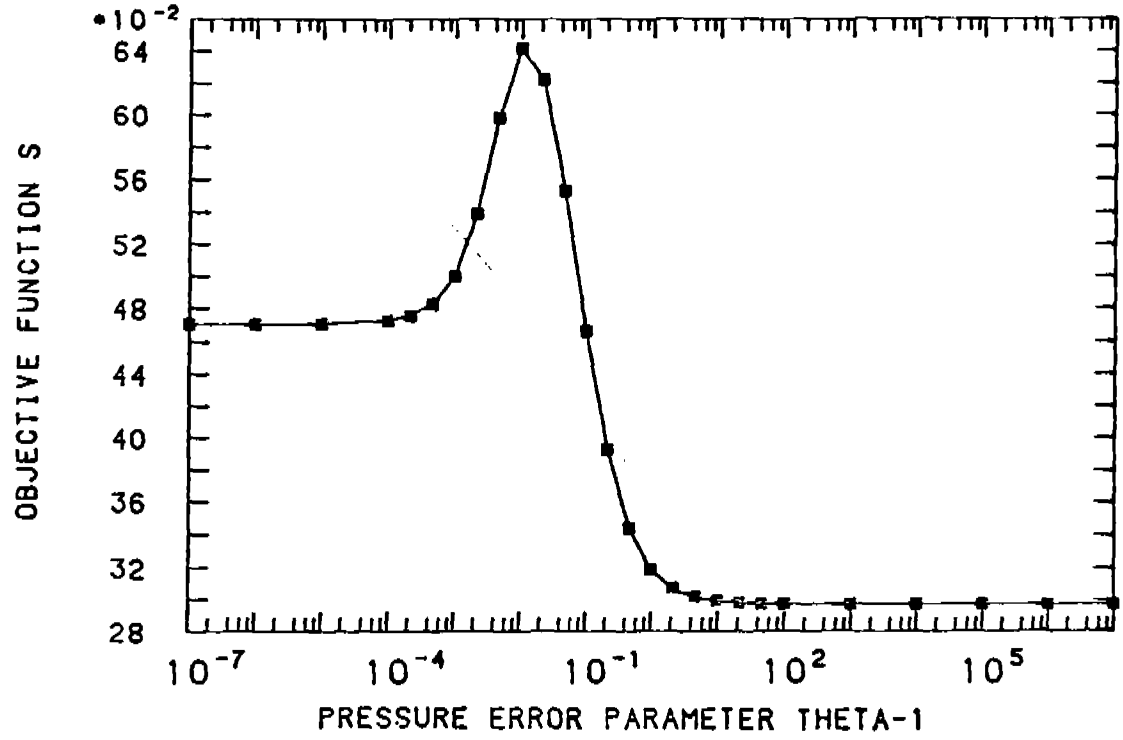


Figure 9. Objective Function S for 1-Tetradecanol Assuming Zero Temperature Error.

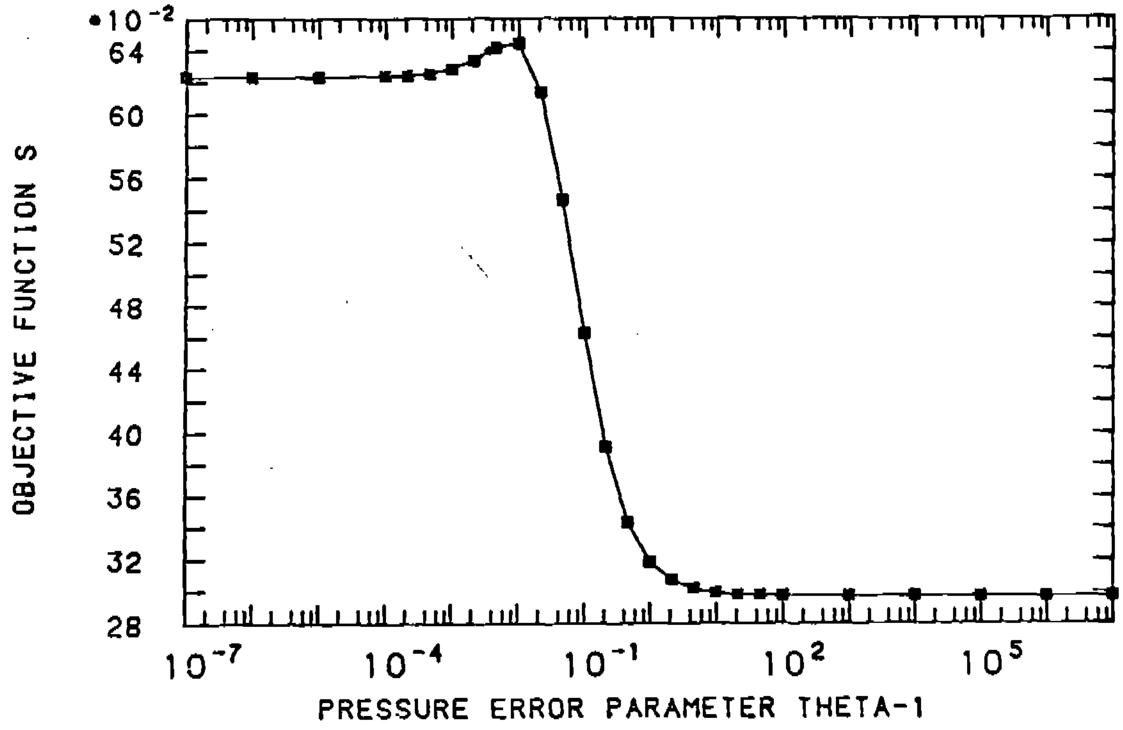


Figure 10. Objective Function S for 1-Tetradecanol Assuming Temperature Error $e_T = 0.1$ K.

standard error (corresponding to an infinite θ_i), as indicated in Table 2. A $S(\theta)$ without a minimum seems to be typical if the number s of data points is small.

The difference between the present results and those by Kemme and Kreps is between one and two standard errors of the Antoine parameters. It is not clear from reference [1] how the Antoine parameters were calculated, but the reported values are between those for constant e_p and for e_p proportional to p. We obtained almost the same Antoine constants as Kemme and Kreps by assuming $e_T = 0$ and

$$e_{pi}/p_R = 0.139 + 0.139\ (p_i + c_{pi})/p_R.$$

The corresponding S was 0.4663 indicating that this is not an optimal choice for the error models.

Table 1. Result from a Simulated Experiment

	Input	Retrieved		
$e_T = e_o\ \theta_2$	0.1 K	(0.0 K)	(0.1 K)	(1.0 K)
e_o	1.0	1.20	1.20	1.11
$e_o\ \theta_1$	0.05	0.056	0.055	0.051
S	---	0.232 944	0.232 945	0.233 230
A	7.0	7.567 ± 0.629	7.567 ± 0.629	7.570 ± 0.631
B	1900 K	2406 ± 551	2406 ± 551	2408 ± 552
C	130 K	181.39 ± 49.88	181.40 ± 49.88	181.51 ± 49.88
c_{AB}	---	0.998 909 68	0.998 909 86	0.998 927 17
c_{AC}	---	0.995 506 28	0.995 506 93	0.995 571 48
c_{BC}	---	0.998 824 72	0.998 824 86	0.998 839 41

The c_{AB}, c_{AC} and c_{BC} are correlation coefficients of the Antoine parameters A, B and C. The temperature error estimate e_T was preset as indicated.

Table 2. Adjustment Results for 1-Tetradecanol

	Present Analysis		Kemme and Kreps[1]
	e_T = 0.1 K	e_T = 0	e_T = 0
Optimal e_p/p_R	0.0294 p/p_R	0.0294 p/p_R	---
S	0.2974	0.2973	---
A	6.2284 ± 0.1846	6.2251 ± 0.1851	6.4840
B (K)	1250.2 ± 106.3	1248 ± 106.7	1412.907
C (K)	76.23 ± 12.02	76.01 ± 12.06	95.368
c_{AB}	0.997 527 0	0.997 525 0	---
c_{AC}	0.991 019 0	0.991 025 3	---
c_{BC}	0.997 877 0	0.997 882 1	---

The c_{AB}, c_{AC}, and c_{BC} are correlation coefficients between the Antoine parameters A, B, and C. The temperature range of observations is between 425.15 and 569.15 K (152 and 296°C).

4. Summary and Conclusion. Multidimensional residuals arise in model fitting problems when more than one component of the observables is subject to adjustment. Because the corrected observations must satisfy constraints representing the model and because such constraints effectively reduce the degrees of freedom for the corrections, the components of multidimenional residual vectors typically are strongly correlated. These correlations must be eliminated or otherwise taken into account when the residuals are analyzed, so that interpretation errors can be avoided. We suggest in this paper to eliminate the model induced correlations by using reduced residuals. The latter generally have less components than the original residuals, and the components are dimensionless. The reduction of the number of components matches the loss of degrees of freedom, so that there are no correlations between components of reduced residuals, except when such correlations are present in the observations. In standard least squares problems (curve fitting in a plane, surface fitting in three dimensions, etc.) the reduced residuals are scalars with an absolute value equal to a norm of the residual and a positive or negative sign, depending on the location of the observation

with respect to the fitted structure. Hence, by introducing reduced residuals one can use in standard least squares problems well known techniques for statistical investigation of scalar quantities. As an example for such application we presented the investigation of heteroscedasticity of vapor pressure measurements. In elementary least squares problems the reduced residuals are identical to the usual weighted residuals.

In conclusion, the introduction of reduced residuals greatly simplifies the investigation of residuals arising in general least squares problems. The concept has a simple geometric interpretation and the routine calculation of reduced residuals easily can be included in least squares utility routines.

REFERENCES

[1] Herbert R. Kemme and Saul I. Kreps, "Vapor Pressure of Primary n-Alkyl Chlorides and Alcohols," Journal of Chemical and Engineering Data, Vol. 14, pp. 98-102, 1969.

[2] Aivars Celmiņš, "A Manual for General Least Squares Model Fitting," US Army Ballistic Research Laboratory Technical Report, ARBRL-TR-02167, June 1979.

[3] Kurt Nielsen, "A Method for Optimizing Relative Weights in Least Squares Analysis," Acta Cristallographica, A33, pp. 1009-1010, 1977.

[4] E. Richard Cohen, ""Extended" Least Squares," Rockwell International Science Center Report, SCTR-76-1, January 1976.

[5] Aivars Celmiņš, "Least Squares Optimization with Implicit Model Equations," in Mathematical Programming with Data Perturbation II, Anthony V. Fiacco, editor, pp. 131-152, Marcel Dekker, New York, 1983.

[6] Aivars Celmiņš, "Least Squares Adjustment with Finite Residuals for Non-linear Constraints and Partially Correlated Data," US Army Ballistic Research Laboratory Report BRL R 1658, July 1973.

[7] Herbit I. Britt and R. H. Luecke, "The Estimation of Parameters in Non-Linear, Implicit Models, Technometrics, Vol. 15, pp. 233-247, 1973.

[8] Allen J. Pope, "Two approaches to Non-Linear Least Squares Adjustements," The Canadian Surveyor, Vol. 28, pp. 663-669, 1974.

[9] William H. Jefferys, "On the Method of Least Squares," The Astronomical Journal, Vol. 85, pp. 177-182, 1980.

APPENDIX
PROPERTIES OF REDUCED RESIDUALS

We provide in this appendix proofs and derivations of the formulas quoted in Section 2.

In the model equations

$$F_i(\xi, t) = 0, \qquad i = 1,\dots, s \tag{A.1}$$

of the minimization problem (2.1) the r-dimensional vector functions $F_i(\xi,t)$ are assumed to be componentwise twice differentiable with respect to all its n + p arguments. We also assume that

$$\text{rank } \partial F/\partial \xi = r \tag{A.2}$$

in a neighborhood of the least squares solution. The condition (A.2) insures that the r equations in each set in Eq. (A.1) are independent. Furthermore, we insure sufficient degrees of freedom for the optimization problem (2.1) by assuming that the following inequalities are satisfied by the dimensions of the problem:

$$0 < p < r \cdot s < n \cdot s \quad . \tag{A.3}$$

The normal equations for the optimization problem can be derived by Lagrange's multiplier technique. To that end we define a modified objective function $\tilde{W}$ by

$$\tilde{W} = \sum_{i=1}^{s} \frac{1}{2} c_i^T R_i^{-1} c_i - \sum_{i=1}^{s} k_i^T F_i(X_i + c_i, t) \quad , \tag{A.4}$$

where the k_i are r-dimensional correlate vectors (Lagrange multipliers). Then the normal equations are obtained by setting equal to zero the partial derivatives of $\tilde{W}$ with respect to the unknown c_i, t and k_i. The result is [6, 7, 8, 9]

$$c_i - R_i F_{xi}^T k_i = 0 \quad , \tag{A.5a}$$

$$\sum_{i=1}^{s} k_i^T F_{ti} = 0 \quad , \tag{A.5b}$$

$$F_i = 0 \quad . \tag{A.5c}$$

Because of the rank condition (A.2) and because R_i is a positive definite matrix, the product $F_{xi}R_iF_{xi}^T$ is also positive definite and one can obtain from Eq. (A.5a)

$$k_i = (F_{xi}\ R_iF_{xi}^T)^{-1}\ F_{xi}\ c_i \quad .$$

Substituting this expression back into Eq. (A.5a) one obtains [6]

$$c_i = R_iF_{xi}^T(F_{xi}R_i\ F_{xi}^T)^{-1}F_{xi}c_i \quad . \tag{A.6}$$

We shall use this important relation later.

In the n-dimensional space of observables we define the inner products

$$(c_i,\ b_i) = c_i{}^T\ R_i^{-1}b_i \tag{A.7}$$

and the elliptic norms

$$||c_i|| = (c_i^T\ R_i^{-1}\ c_i)^{1/2} \quad . \tag{A.8}$$

These definitions are possible because the R_i are positive definite. We notice that the norm (A.8) is dimensionless.

The model equation $F_i(\xi,t) = 0$ defines a (n - r)-dimensional hypersurface in the n-dimensional space of observables. Let $X_i + c_i$ be a point of that surface. The hyperplane orthogonal to the surface $F_i = 0$ at that point is r-dimensional and is spanned by the rows of the matrix

$$M_{Fi}^T = F_{xi}\ R_i \quad , \tag{A.9}$$

that is, for an arbitrary r-vector $\tilde{a}_i$ the corresponding n-vector

$$n_i = M_{Fi}\ \tilde{a}_i = R_i\ F_{xi}^T\tilde{a}_i \tag{A.10}$$

is orthogonal to $F_i = 0$, whereby the orthogonality is defined in terms of the inner product (A.7). To show this, we compute the inner product of n_i with a vector b_i tangential to the surface. The latter vector satisfies the equation

$$F_{xi} b_i = 0 \quad . \tag{A.11}$$

The inner product is

$$\begin{aligned}(n_i, b_i) = (M_{Fi} \tilde{a}_i, b_i) &= \tilde{a}_i^T M_{Fi}^T R_i^{-1} b_i \\ &= \tilde{a}_i^T F_{xi} b_i = 0\end{aligned} \tag{A.12}$$

for arbitrary $\tilde{a}_i$.

Next, we normalize the matrix M_{Fi} by defining

$$\begin{aligned}N_{Fi}^T &= (F_{xi} R_i F_{xi}^T)^{-1/2} M_{Fi}^T \\ &= (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi} R_i \quad .\end{aligned} \tag{A.13}$$

The r rows of the matrix N_{Fi}^T again span the orthogonal hyperplane, but they are unit vectors orthogonal to each other. We show this by computing the inner product of N_{Fi} with itself:

$$\begin{aligned}(N_{Fi}, N_{Fi}) &= N_{Fi}^T R_i^{-1} N_{Fi} \\ &= (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi} R_i F_{xi}^T (F_{xi} R_i F_{xi}^T)^{-1/2} = I.\end{aligned} \tag{A.14}$$

That N_{Fi} spans the same subspace as M_{Fi} follows from the formula

$$N_{Fi} \tilde{b}_i = R_i F_{xi}^T (F_{xi} R_i F_{xi}^T)^{-1/2} \tilde{b}_i = M_{Fi} \tilde{a}_i \tag{A.15}$$

for arbitrary $\tilde{b}_i$ and a corresponding proper $\tilde{a}_i$.

Next we show that the least squares residual c_i is an element of the orthogonal space. According to Eq. (A.6) the residual satisfies the relation

$$c_i = R_i F_{xi}^T (F_{xi} R_i F_{xi}^T)^{-1/2} (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi} c_i \tag{A.16}$$

$$= N_{Fi} N_{Fi}{}^T R_i{}^{-1} c_i = N_{Fi} \breve{b}_i$$

with a proper $\breve{b}_i$ and, therefore, is an element of the subspace, as claimed.

The reduced residual a_i is defined by Eq. (2.2), or

$$a_i = (F_{xi} R_i F_{xi}^T)^{-1/2} F_{xi} c_i = N_{Fi}^T R_i^{-1} c_i \quad . \tag{A.17}$$

It is an r-vector with dimensionless components, and it represents the residual c_i in the orthogonal hyperplane. According to Eq. (A.16) the relation (A.17) also can be solved for the c_i:

$$c_i = N_{Fi} a_i \quad . \tag{A.18}$$

Hence, either of the two vectors a_i and c_i are uniquely determined by the other.

The Euclidean norm of the reduced residual is, because of Eqs. (A.17) and (A.13),

$$||a_i|| = (a_i^T a_i)^{1/2} = \{c_i^T [F_{xi}^T (F_{xi} R_i^T F_{xi})^{-1} F_{xi} c_i]\}^{1/2} \tag{A.19}$$

$$= (c_i{}^T R_i^{-1} c_i)^{1/2} = ||c_i||$$

Eq. (A.19) shows that the Euclidean norm of the r-vector a_i is equal to the elliptic norm (A.8) of the n-vector c_i.

A relation similar to (A.17) also exists between the reduced residual a_i and the unknown true observational error $\overset{*}{c}_i$ of the observation X_i. Let $\overset{*}{t}$ be the true value of the model parameter. Then by definition

$$F_i(X_i + \overset{*}{c}_i, \overset{*}{t}) = 0 \quad . \tag{A.20}$$

Expanding the model function F_i at the least squares position one obtains

$$F_i(X_i + \overset{*}{c}_i, \overset{*}{t}) = F_i(X_i + c_i, t) + F_{xi}(\overset{*}{c}_i - c_i) + F_{ti}(\overset{*}{t} - t) + \ldots \quad (A.21)$$

The left hand side and the first term of the right hand side are zero. If one neglects terms of higher order in $\overset{*}{c}_i - c$ and $\overset{*}{t} - t$, then Eq. (A.21) yields the following relation between c_i and $\overset{*}{c}_i$

$$F_{xi}\, c_i = F_{xi}\, \overset{*}{c}_i + F_{ti}\, (\overset{*}{t} - t) \quad . \quad (A.22)$$

Eq. (A.22) is a relation between projections of c_i, $\overset{*}{c}_i$ and $\overset{*}{t} - t$ on the orthogonal subspace. Because c_i is an element of that subspace one can use Eq. (A.17) and calculate the contributions of the other two projections to the reduced residual a_i:

$$a_i = (F_{xi}\, R_i\, F_{xi}^T)^{-1/2}\, F_{xi}\, \overset{*}{c}_i + (F_{xi}\, R_i\, F_{xi}{}^T)^{-1/2}\, F_{ti}(\overset{*}{t} - t) \quad . \quad (A.23)$$

Figure 1 illustrates the relation (A.23) in a planar curve fitting case.

The relation (A.17) has the inversion (A.18) that allows one to calculate c_i if a_i is given. A similar inversion of Eq. (A.23) would permit one to compute the unknown true errors $\overset{*}{c}_i$ in terms of the a_i that are known from the least squares fitting. However, such an inversion of Eq. (A.23) is generally not possible. For Eq. (A.17) it was derived by using Eq. (A.6) which is satisfied by the least squares residuals c_i, but not by the true errors $\overset{*}{c}_i$. This result is geometrically obvious because only a projection of $\overset{*}{c}_i$ enters Eq. (A.23) and one cannot reconstruct a vector from its projection. An exception is the special case r = n, that is, the case where the constraint equation $F_i = 0$ has as many components as the space of observables. In that case the Jacobian matrix F_{xi} can be inverted and one obtains from Eq. (A.22), for instance, the expression

$$\overset{*}{c}_i = c_i - F_{xi}^{-1}\, F_{ti}(t - \overset{*}{t}) \quad (A.24)$$

as an estimate of the true errors. A corresponding relation between c_i^* and a_i, which in this case also has n components, can be derived by substituting Eq. (A.18) into Eq. (A.24). However, in this special case the reduced residuals offer the only advantage that they have dimensionless components. Therefore, an investigation of heteroscedasticity can be more effectively done in this case in terms of the c_i instead of the reduced residuals a_i, because of the simplicity of the relation (A.24) between c_i and c_i^*.

A least squares model fitting can determine only the projection of the true observational errors on the space orthogonal to the surface F = 0. Because other components of the true errors are lost it is reasonable to carry out any investigations of the residuals only in the orthogonal space. This is done by using the reduced residuals, which contain all the information about the true errors that is present in the least squares results.

APPLICATION OF THE PRINCIPAL COMPONENT METHOD TO TRAJECTORY ESTIMATION

William S. Agee and Robert H. Turner
Mathematical Services Branch
Data Sciences Division
National Range Operations Directorate
US Army White Sands Missile Range
White Sands Missile Range, New Mexico 88002

1. TRAJECTORY ESTIMATION. Measurements of range, azimuth, and elevation from several different radars are used to estimate the cartesian position coordinates of a vehicle trajectory at a sequence of times, t_i, i=1,N which cover the entire trajectory. Since the measurements are subject to systematic errors as well as random measurement errors, we also want to estimate the systematic error parameters or biases in addition to the trajectory coordinates. The resulting estimation problem is a combined linear and nonlinear estimation problem in which the trajectory coordinates appear as nonlinear parameters in the measurements and the biases appear as linear parameters in the measurements.

Let $h_\alpha(\bar{x}_i)$ be a measurement function where $\vec{x}_i$ is the cartesian position vector to the trajectory at time t_i. If we have M different radars observing the trajectory, then α = 1, 3M. For a range measurement from the $p^{\underline{th}}$ radar

$$h_\alpha(\bar{x}_i) = [(x_i-x_p)^2 + (y_i-y_p)^2 + (z_i-z_p)^2]^{1/2} \qquad (1)$$

where (x_p, y_p, z_p) are the cartesian coordinates of the origin of the local cartesian coordinate system at the $p^{\underline{th}}$ radar. For an azimuth measurement from the $p^{\underline{th}}$ radar the measurement function is,

$$h_\alpha(\bar{x}_i) = \tan^{-1} \frac{x_i - x_p}{y_i - y_p} \qquad (2)$$

For an elevation measurement from the $p^{\underline{th}}$ radar

$$h_\alpha(\bar{x}_i) = \tan^{-1} \frac{z_i - z_p}{[(x_i-x_p)^2 + (y_i-y_p)^2]^{1/2}} \qquad (3)$$

Let $z_\alpha(t_i)$ denote the observed value of the $\alpha^{\underline{th}}$ measurement. The observations are modeled as,

$$z_\alpha(t_i) = h_\alpha(\bar{x}_i) + b_\alpha + e_\alpha(i) \qquad (4)$$

where b_α is a constant measurement bias and $e_\alpha(i)$ is a zero mean, random measurement error. Let b be a 3M-dimensional bias vector $b^T = [b_1\ b_2 -- b_{3M}]$. Then the measurement model can be represented as

$$z_\alpha(t_i) = h_\alpha(\bar{x}_i) + s_\alpha b + e_\alpha(i) \tag{5}$$

where s_α is a row vector with a one in the $\alpha^{\underline{th}}$ entry and zeros in all other entries.

$$s_\alpha = [0\ \ 0\ ---0\ \underset{\alpha^{\underline{th}}\ \text{position}}{\overset{1}{\uparrow}}\ 0\ ---0] \tag{6}$$

Let $R_\alpha(t_i)$ be known variances of the random measurement errors, $e_\alpha(\,i\,)$.

The estimation problem to be considered is to minimize,

$$\sum_{i=1}^{N}\sum_{\alpha=1}^{3M}(z_\alpha(t_i)-h_\alpha(\bar{x}_i)-S_\alpha b)^2\ R^{-1}_\alpha(t_i) \tag{7}$$

with respect to $\bar{x}_i$, i = 1, N and b. Differentiating (7) with respect to $\bar{x}_i$ and b results in the nonlinear normal equations

$$\sum_{\alpha=1}^{3M} H^T_\alpha(\hat{x}_i)R^{-1}_\alpha(t_i)(z_\alpha(t_i)-h_\alpha(\hat{x}_i)-S_\alpha\hat{b}) = 0 \quad i = 1,N \tag{8}$$

$$\sum_{i=1}^{N}\sum_{\alpha=1}^{3M} S^T_\alpha R^{-1}_\alpha(t_i)(z_\alpha(t_i)-h_\alpha(\hat{x}_i)-S_\alpha\hat{b}) = 0 \tag{9}$$

where $\hat{x}_i$ is the estimate of $\bar{x}_i$ and $\hat{b}$ is the estimate of b. In (8) $H_\alpha(\bar{x}_i)$ is the derivative, $\dfrac{\partial h_\alpha(\bar{x}_i)}{\partial \bar{x}_i}$. In order to solve the normal equations, they are linearized about a guess trajectory, $x_i^{(s)}$. Let $x_i^{(s)}$, i = 1, N and $b^{(s)}$ satisfy (8), i.e.,

$$\sum_{\alpha=1}^{3M} H_\alpha{}^T(x_i{}^{(s)})R^{-1}_\alpha(t_i)(z_\alpha(t_i)-h_\alpha(x_i{}^{(s)})-S_\alpha b^{(s)}) = 0 \quad i = 1, N \tag{10}$$

If (8) is linearized about $x_i^{(s)}$ and $b^{(s)}$, we obtain

$$(\hat{x}_i - x_i^{(s)}) = -A_i^{-1} A_{i,N+1}^{T} (\hat{b} - b^{(s)}) \tag{11}$$

where

$$A_i = \sum_{\alpha=1}^{3M} H_\alpha^{T}(x_i^{(s)}) R_\alpha^{-1}(t_i) H_\alpha(x_i^{(s)}) \tag{12}$$

and

$$A_{i,N+1}^{T} = \sum_{\alpha=1}^{3M} S_\alpha^{T} R_\alpha^{-1}(t_i) H_\alpha(x_i^{(s)}) \tag{13}$$

A_i is 3 X 3 and $A_{i,N+1}$ is 3 X 3M. Linearizing the second normal equation, (9), about $x_i^{(s)}$ and solving for $\hat{b}$ gives the result,

$$\hat{b} - b^{(s)} = \left(\sum_{i=1}^{N} \sum_{\alpha=1}^{3M} S_\alpha^{T} R_\alpha^{-1}(t_i) S_\alpha - \sum_{i=1}^{N} A_{i,N+1}^{T} A_i^{-1} A_{i,N+1} \right)^{-1} \left[\sum_{i=1}^{N} \sum_{\alpha=1}^{3M} S_\alpha^{T} R_\alpha^{-1}(t_i) (z_\alpha(t_i) - h_\alpha(x_i^{(s)}) - S_\alpha b^{(s)}) \right] \tag{14}$$

(11) and (14) for $\hat{x}_i - x^{(s)}$ and $\hat{b} - b^{(s)}$ are the basic equations for trajectory estimation. The solution to the normal equations are obtained by successive relinearization and solution of (8) and (9).

2. APPLICATION OF PRINCIPAL COMPONENTS REGRESSION. Although there are no convergence problems in solving the normal equations iteratively for the M-station radar case, another problem which is fairly common in the solution of linear least squares problems also occurs frequently in trajectory estimation. Very often, the estimate of the bias vector, b, converges to a solution for which several of the components are too large and may have the wrong sign. Sometimes the bias estimate is obviously erroneous. One obviously erroneous case which arises frequently is that the elevation bias components will all be large and of the same sign. This problem of the estimated bias vector being too long is usually attributed to multicollinearity among the predictor variables in the linear least squares problem. The problem in the linear estimation case is often successfully treated by some method of biased estimation. The problem has not been properly recognized or successfully treated when it arises in trajectory estimation. Although the existence of these erroneous bias estimates has been

recognized in trajectory estimation, the source of the difficulty was not properly recognized. Some workers in trajectory estimation have stated that the existence of this problem demonstrates the need to specify a prior distribution for the biases in order to "tie down" or statistically constrain the bias estimates. It does not take much experience in using these priors for trajectory estimation to realize that the problem of inflated bias estimates is as much present with the prior as without the prior. We have attempted to treat this problem both with ridge regression and with principal components regression. We have had some success with both methods but neither method has entirely solved the problem. Of the two methods we have had the most success with principal components, probably because it is easier to apply. The principal components method also illuminates a problem that arises in the application of either method.

The linearized equation (14) to be solved for the bias estimate can be written as

$$Q\hat{b} = U$$

where

$$Q = \sum_{i=1}^{N}\sum_{\alpha=1}^{3M} S_{\alpha}^{T} R_{\alpha}^{-1}(t_i) S_{\alpha} - \sum_{i=1}^{N} A_{i,N+1}^{T} A_{i}^{-1} A_{i,N+1} \tag{15}$$

and

$$U = \sum_{i=1}^{N}\sum_{\alpha=1}^{3M} S_{\alpha}^{T} R_{\alpha}(t_i)(Z_{\alpha}(t_i) - h_{\alpha}(x_i^{(s)}) - S_{\alpha} b^{(s)}) + Qb^{(s)} \tag{16}$$

Partition the bias vector into angle components b_a and range components b_r so that the linear bias estimation equation is

$$\begin{bmatrix} Q_a & R \\ R^T & Q_r \end{bmatrix} \begin{bmatrix} \hat{b}_a \\ \hat{b}_r \end{bmatrix} = \begin{bmatrix} U_a \\ U_r \end{bmatrix} \tag{17}$$

Scale the bias estimates according to

$$\begin{bmatrix} \hat{b}_a \\ \hat{b}_r \end{bmatrix} = \begin{bmatrix} D_a^{-1} & 0 \\ 0 & D_r^{-1} \end{bmatrix} \begin{bmatrix} \hat{b}'_a \\ \hat{b}'_r \end{bmatrix} \tag{18}$$

Where D_a and D_r are diagonal matrices chosen so that $D_a^{-1}Q_aD_a^{-1} = Q_a'$ and $D_r^{-1}Q_rD_r^{-1} = Q_r'$ each have diagonals of unity. The transformed bias estimation equation is

$$\begin{bmatrix} Q_a' & R' \\ R'^T & Q_r' \end{bmatrix} \begin{bmatrix} \hat{b}_a' \\ \hat{b}_r' \end{bmatrix} = \begin{bmatrix} U_a' \\ U_r' \end{bmatrix} \qquad (19)$$

with $R' = D_a^{-1}RD_r^{-1}$ and $U_a' = D_a^{-1}U_a, U_r^{-1} = D_r^{-1}U_r$. Suppose the bias vector is further transformed by

$$\begin{bmatrix} \hat{b}_a \\ \hat{b}_T' \end{bmatrix} = T\beta \qquad (20)$$

where T is orthogonal and diagonalizes

$$Q' = \begin{bmatrix} Q_a' & R' \\ R'^T & Q_r' \end{bmatrix}. \qquad (21)$$

In principal components regression the components of β which correspond to very small eigenvalues of Q' are set to zero. Specifically, let r be the largest integer for which

$$\frac{\sum\limits_{i=3M-r}^{3M} \gamma_i}{3M} \leq 10^{-2} , \qquad (22)$$

where γ_i i=1, 3M are the eigenvalues of Q' and we have ordered the γ_i so that $\gamma_{i+1} \leq \gamma_i$. Define $\beta_i^* = 0$ for $r \leq i \leq 3M$ and $\beta_i^* = \beta_i$ otherwise. The principal components solution of the original system of equations, denoted by $[b_a^* \; b_r^*]$, is defined by

$$\begin{bmatrix} b_a^* \\ b_r^* \end{bmatrix} = \begin{bmatrix} D_a^{-1} & 0 \\ 0 & D_r^{-1} \end{bmatrix} T\beta^* \qquad (23)$$

Consider the following example from WSMR data. We have three radars, R122, R123, and R395 tracking a level flying target, flying at an altitude of about 30,000 ft. The graph of Fig 1 shows the relative geometry of the target trajectory and radars. The least square estimates of the radar biases obtained for this example are

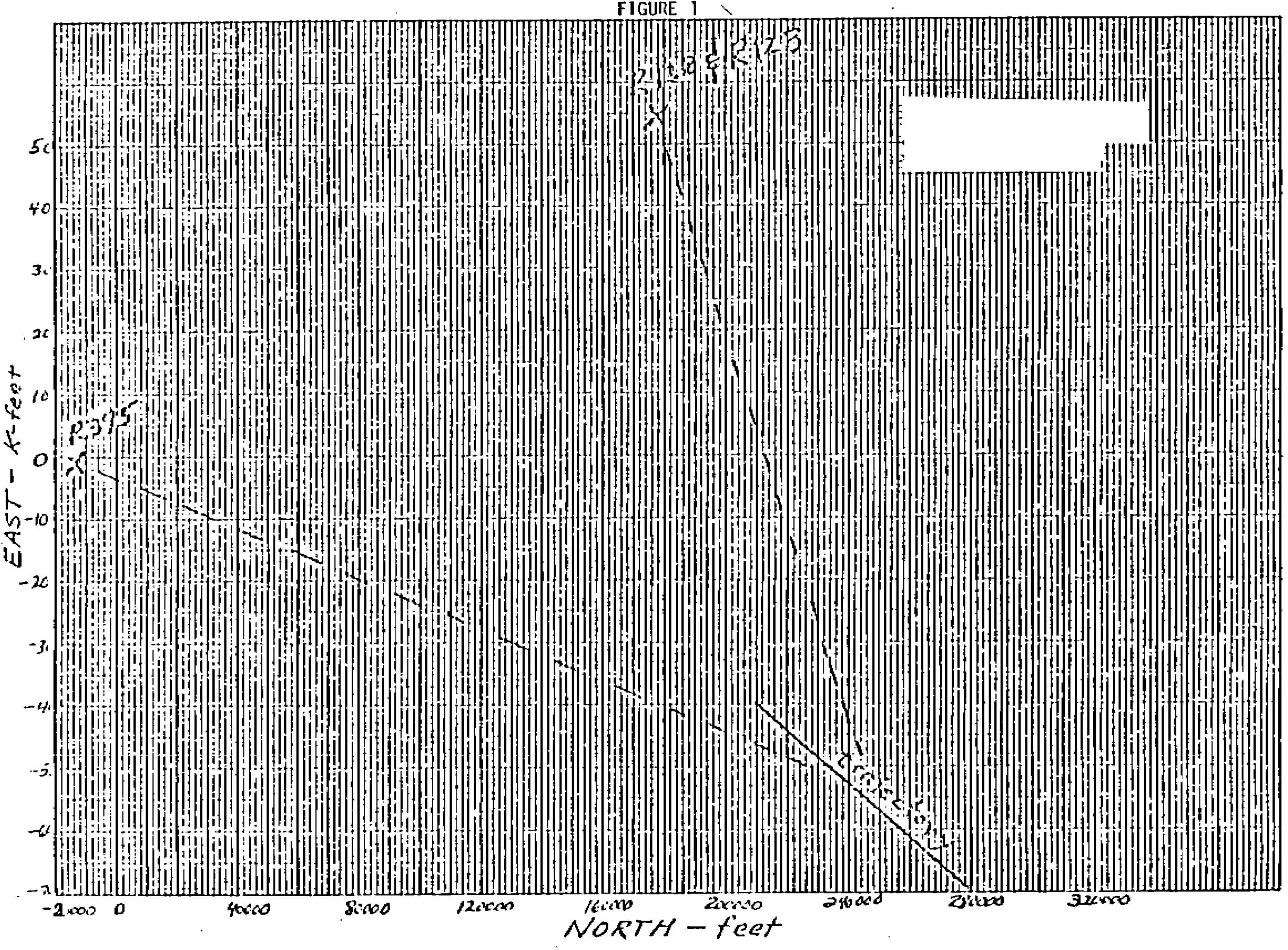
	R122	R123	R395
Range bias (ft)	118.3	114.8	63.7
Azimuth Bias (miliradians)	.116	.058	.148
Elevation bias (miliradians	-.737	-.947	-.538

The values of the elevation bias estimates, which are all large and negative illustrate a commonly occuring type of erroneous solution in radar trajectory estimation. In this example we are able to confirm that the radar bias estimates above are greatly in error. Using measurements from tracking cameras we are able to obtain trajectory estimates which are much more accurate than trajectory estimates obtained from radars. By comparing the optically derived trajectory against the radar measurements we obtain the following radar measurement bias estimates, which we use as a standard.

	R122	R123	R395
Range bias (ft)	157.3	152.9	80.3
Azimuth bias (miliradians)	.05	.02	.09
Elevation bias (miliradians)	.11	-.08	-.09

The large errors in the radar bias estimates especially in elevation are readily apparent.

FIGURE 1


21-28 & R125
P375
Crossing
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
EAST - K-feet
-2000
0
4000
8000
12000
16000
20000
24000
28000
32000
NORTH - feet

The eigenvalues for the Q' matrix and the corresponding components of the solution vector $\hat{\beta}$ are:

γ_i	β_i
1.9306182	-61.638411
1.7776564	27.106658
1.7112154	17.060269
1.28B728	-25.130580
1.2173482	25.571920
1.0753069	88.890445
$.45750007x10^{-2}$	-481.67132
$.16010719x10^{-2}$	-83.385294
$.30452861x10^{-3}$	55.338603

Using the criterion stated in (22) principal components regression will set the last three components of β^* to zero. The principal component solution for the radar bias estimates is

	R122	R123	R395
Range bias (ft)	11.9	8.2	-22.3
Azimuth bias (miliradians)	.09	.03	.47
Elevation bias (miliradians)	.13	-.08	-.13

The principal components method for this example gives large errors in the estimates of the range biases and a large error in the azimuth bias estimate for R395. Thus, it appears that the principal components method does not yield useful results when applied to this example. However, we have found that the principal components method does give good results when applied to this example in a slightly different way.

Consider the partitioned bias estimation equation in (19). Transform the angle bias vector $\hat{b}_a$ as $\hat{b}'_a = T\hat{\beta}$, where T is orthogonal. The range bias, $\hat{b}'_r$, can be eliminated from (19) by substitution. We obtain

$$\hat{b}'_r = Q'^{-1}_r(U'_r - R'^T T\hat{\beta})$$

and

$$T^T(Q_a' - R'^T Q_r'^{-1} R')T\hat{\beta} = T^T(U_a' - R'^T Q_r'^{-1} U_r') \tag{24}$$

Choose T so that

$$T^T(Q_a' - R'^T Q_r'^{-1} R')T = \Gamma \tag{25}$$

Then $\Gamma\hat{\beta} = T^T(U_a' - R'^T Q_r'^{-1} U_r')$ (26)

The values of the radar bias estimates for the current example obtained from applying the principal component method to this reduced problem are:

	R122	R123	R395
Range bias (ft)	147.7	144.1	70.2
Azimuth bias (miliradians)	.01	-.04	.11
Elevation bias (miliradians)	.13	-.08	-.06

The biases computed by applying the principal component method to the reduced bias estimation problem are statistically compatible with the radar biases derived by comparison with optical tracking data. Another example of inflated radar bias estimates comes from a recent missile vs. drone engagement at WSMR. This example will indicate some additional difficulties in applying principal components to the radar bias estimate problem. In this example we have three radars, R124, R125, and R442 tracking the missile. The least squares estimation of trajectory and radar biases obtained the following bias estimates.

	R124	R125	R442
Range bias (ft)	253	307	164
Azimuth bias (miliradians)	.21	.21	-.57
Elevation bias (miliradians)	-.39	-.71	-.85

For this example the eigenvalues of the reduced matrix, $Q'_a - R'^T Q'^{-1}_r R'$, and the corresponding components of the solution vector $\hat{\beta}$ are:

γ_i	$\hat{\beta}_i$
1.8357080	38.432177
1.4980308	-42.134995
1.0704788	7.2953978
1.2442094	26.848291
.25488164	-171.70691
.096690436	-442.92985

If we apply the criterion stated in (22) for zeroing components of $\hat{\beta}$ corresponding to small eigenvalues, we find that no components should be zeroed since even the two smallest eigenvalues are not very small. Nevertheless, we zero the components of β corresponding to the two smallest eigenvalues. This results in the bias estimates,

	R124	R125	R442
Range bias (ft)	263	308	193
Azimuth bias (miliradians)	-.07	-.05	-.12
Elevation bias (miliradians)	.05	-.08	0

From optical tracking data on this drone we computed the following estimates of the radar biases.

	R124	R125	R442
Range bias (ft)	254	297	204
Azimuth bias (miliradians)	0	-.04	-.16
Elevation bias (miliradians)	0	-.08	-.09

The bias estimates obtained by applying principal components are much more compatible with the optically derived bias estimates than the bias estimates obtained from least squares. We have tried several other examples from which we obtain similar results and conclusions.

We believe that we have made significant progress in solving the problem of inflated measurement bias estimates in trajectory estimation. However, before we could be confident in the routine application of the principal component method described above to M-station radar trajectory estimation, several questions should be carefully considered.

(1) Why must the principal component method be applied to the reduced bias estimation problem (angles only) rather than the full bias estimation problem?

(2) What criterion should be used zeroing components of $\hat{\beta}$ corresponding to small eigenvalues?

(3) Should the principal component method somehow be applied to obtaining improved range bias estimates?

(4) Are there other biased estimation methods such as ridge regression, fractional rank regression, etc., which would yield better results than the principal component method when applied to the estimation of measurement biases in M-station radar trajectory estimation?

OPTIMAL REDUCED ORDER CONTROLLERS
for
DISCRETE-TIME LINEAR SYSTEMS
with
PARTIAL STATE ESTIMATION

Maurice F. Hutton

IBM Corporation
Federal Systems Division
Owego, New York 13827

ABSTRACT

The solution to the standard steady state LQG problem is a controller whose order equals the order of the plant model. For many high order plant models, a reduced order controller which is easier to mechanize can be used with good performance. A methodology is described for solving the discrete-time, steady state LQG problem with the constraints that the controller be of reduced order and that linear combinations of the controller states estimate particular plant model state variables. Formulas for simplifying the gradient calculations are introduced.

NOMENCLATURE

Dimensions

Symbol		Definition
k	=	number of plant states
l	=	number of plant inputs
m	=	number of plant outputs
p	=	number of plant noise inputs
r	=	number of controller states
s	=	number of variables being tracked

Matrices

Size	Symbol		Definition
k x k	A	=	plant system coefficients
k x l	B	=	plant input coefficients
m x k	C	=	plant output coefficients
k x p	D	=	plant input noise coefficients
s x s	E	=	performance weights on tracking error
r x r	F	=	controller system coefficients
r x m	G	=	controller input coefficients
l x r	H	=	controller output coefficients

s x k	K =	tracking coefficients
s x r	L =	approximate tracking coefficients
k+r x k+r	M =	closed loop system coefficients
k+r x p+m	N =	closed loop input noise coefficients
k+r x k+r	P =	variance of closed loop state vector
k x k	Q =	performance weights on plant state
l x l	R =	performance weights on plant input
p+m x p+m	S =	variance of closed loop input noise vector
s x k+r	T =	closed loop tracking coefficients
k+r x k+r	U =	composite performance weights
p x p	V =	variance of plant input noise vector
m x m	W =	variance of plant measurement noise vector
k+r x k+r	Y =	Lagrange multipliers
k+r x k+r	Z =	performance weights on closed loop state

Vectors

Size	Symbol	Definition
s	e =	tracking error
k+r	f =	closed loop state
p+m	g =	closed loop input noise
r	q =	controller state
l	u =	plant input
p	v =	plant input noise
m	w =	plant measurement noise
k	x =	plant state
m	y =	plant output
s	z =	variables being tracked
s	h =	estimates of the variables being tracked

Scalars

Symbol	Definition
b =	step size in steepest descent method
J =	performance index
$\bar{J}$ =	modified performance index
n =	discrete-time index

INTRODUCTON

The motivation for this study was an idea for improving the design used to fly target aircraft in formation under computer control. This Drone Formation Control System (DFCS), currently in operation at the White Sands Missle Range in New Mexico, places a premium on the complexity of the control design because the computations must be performed in real time. However, the plant models for a single drone which include airframe, actuator, and autopilot dynamics have had as many as 32 states. Applying the standard steady state LQG method in such a case would result in a 32 state feedback controller.

An attractive alternative is to design a reduced order controller that minimizes a quadratic performance index. The minimum value of the performance index for the reduced order design will be larger than the value achieved by the high order design. Nevertheless, a reduced order drone controller should meet the performance requirements since only a few of the 32 modes are dominant. The decrease in complexity is expected to be worth the small difference in performance. This property applies to many other applications involving high order linear models.

In the reduced order approach the designer specifies the order and structure of the controller. Hence the design problem is to find the controller gains that minimize the performance index. The theory for solving this problem in the continuous-time case is described in [1]-[4].

A disadvantage of using the reduced order approach is that the controller states lose their physical meaning. For the standard LQG method, the controller states (for a particular realization) are estimates of the plant states. This is not true for the reduced order controller. Because DFCS must not only control but also track the drones, this limitation of the reduced order approach presents a problem. Fortunately, estimates of all 32 drone states are not needed. For tracking purposes, the DFCS controller only needs to estimate position and velocity.

To resolve the tracking problem, the reduced order approach is extended to provide estimates of some state variables of special interest to the designer. To be more exact, linear combinations of the state variables may be estimated. The extension is accompished by modifying the performance index and the structure of the controller.

The basic equations for designing the reduced order controller with partial state estimation are described below. The definitions of the symbols used in the equations are given in the nomenclature section.

PROBLEM FORMULATION

The linear, time-invariant model of the plant to be controlled is

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n) + Dv(n) \\ y(n) &= Cx(n) + w(n) \end{aligned} \tag{1}$$

In this state space representation the vectors x, u, y denote the state, input, and output, respectively. The random vectors v and w modeling the plant disturbances and measurement errors are zero mean white noise sequences with variance matrices V and W, respectively.

$$\bar{E}\{ vv' \} = V \qquad \bar{E}\{ ww' \} = W \tag{2}$$

The operator $\bar{E}\{ \}$ is the steady state expectation.

The feedback controller is assumed to have the following structure

$$q(n+1) = Fq(n) + Gy(n) \qquad u(n) = -Hq(n) \tag{3}$$

The order of the controller r is selected by the designer and is less than or equal to the order k of the plant model.

$$1 \leq r \leq k \tag{4}$$

In addition to computing the feedback control, the controller must also compute estimates h of particular plant model variables z which are linear combinations of the plant states. The estimates h are linear combinations of the controller states.

$$z(n) = Kx(n) \tag{5}$$

$$h(n) = Lq(n) \tag{6}$$

The tracking error e is the difference between the estimated and actual values for the particular plant model variables to be tracked.

$$e(n) = h(n) - z(n) \tag{7}$$

The performance index is the steady state expectation of a quadratic form weighting the state, input, and tracking error.

$$J = \bar{E}\{ x'Qx + u'Ru + e'Ee \} \tag{8}$$

The first 2 terms in (8) correspond to the standard definition of the performance index for the steady state LQG discrete-time problem. The additional term provides a means of extending the controller design to perform partial state estimation.

To summarize, the objective is to find the matrices F, G, H, L that minimize the performace index (8) subject to conditions (1)-(7). A block diagram of the closed loop structure is shown in Figure 1. It is assumed that the design parameters or controller gains in the minimization are particular elements of the matrices F, G, H, L. This assumption is more general than necessary but adequate for many applications, including the applications of current interest to the author. The designer specifies which matrix elements are to be design parameters as part of the problem setup. In order to have a unique solution, not all the elements of the controller coefficient matrices can be considered as design parameters. Denery [5] has shown that only r(l+m) unknown parameters are needed in general to determine the matrices F, G, H defining the feedback control.

It is also assumed that a solution exists. The conditions guaranteeing that the optimization procedure will converge to a unique solution remain to be studied.

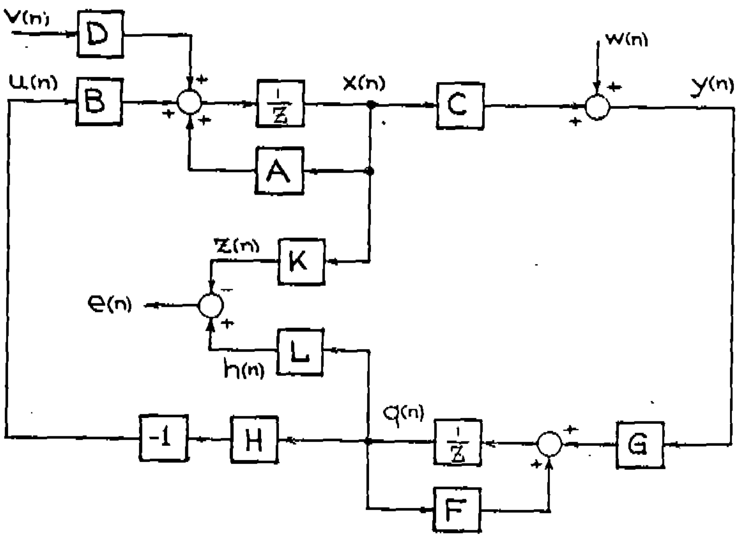


Figure 1 Block Diagram of Closed Loop System with Reduced Order Controller

THEORY

The first phase of the derivation is to express the closed loop dynamics in state space form.
Let

$$f = \left[\begin{array}{c} x \\ \hline q \end{array}\right] \qquad g = \left[\begin{array}{c} v \\ \hline w \end{array}\right] \tag{9}$$

$$M = \left[\begin{array}{c|c} A & -BH \\ \hline GC & F \end{array}\right] \qquad N = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & G \end{array}\right] \tag{10}$$

Then the closed loop dynamics are described by

$$f(n+1) = Mf(n) + Ng(n) \tag{11}$$

Also let

$$T = \left[\begin{array}{c|c} -K & L \end{array}\right] \tag{12}$$

Then the tracking error in terms of the closed loop dynamics is

$$e(n) = Tf(n) \tag{13}$$

The second phase is to find an equation for the variance P of the closed loop state vector f as a function of the variance S of the closed loop input noise vector g .
By definition

$$\begin{aligned} \bar{E}\{ ff' \} &= P \\ \bar{E}\{ gg' \} &= S \end{aligned} \tag{14}$$

which from (2) and (9) imply that

$$S = \left[\begin{array}{c|c} V & 0 \\ \hline 0 & W \end{array}\right] \tag{15}$$

Since (11) is a linear, time-invariant discrete-time system excited by a white noise seqence

$$P = MPM' + NSN' \tag{16}$$

The third phase is to express the performance index in a form more suitable for differentiation. Rewriting (8) gives

$$J = \bar{E}\{ \mathrm{tr}[Qxx' + Ruu' + Eee'] \} \qquad (17)$$

The operator tr[ ] denotes the trace of a square matrix which equals the sum of the diagonal elements. The performance index is next written in terms of the closed loop state vector. In order to do this, define

$$Z = \left[\begin{array}{c|c} Q & 0 \\ \hline 0 & H'R\,H \end{array} \right] \qquad (18)$$

Substituting (3) in to the 2nd term and (7) into the 3rd term of (17) gives after applying (5), (6), (9), (12), and (13)

$$J = \bar{E}\{ \mathrm{tr}[Zff' + ETff'T'] \} \qquad (19)$$

Using the commutive property of the trace operator, (19) simplifies to

$$J = \bar{E}\{ \mathrm{tr}[Uff'] \} \qquad (20)$$

where

$$U = Z + T'ET \qquad (21)$$

Transposing the trace and steady state expectation operators plus substituting (14) results in

$$J = \mathrm{tr}[UP] = \mathrm{tr}[PU] \qquad (22)$$

The problem can now be stated as finding the matrices F, G, H, L that minimize (22) subject to the equality constraints imposed by (16).

This problem is equivalent to minimizing the following function in which the independent variables are the unknown elements in the matrices F, G, H, L, P, Y:

$$\bar{J} = \mathrm{tr}[UP + (MPM' + NSN' - P) Y] \qquad (23)$$

The matrix Y contains the Lagrange multipliers. The use of the modified performance index (23) instead of (22) and (16) removes the constraint equations. The minimal solution of (23), which is also the minimal solution of (22), is computed numerically as described in the next section.

COMPUTATIONS

The optimum solution of (23) can be found by setting to zero the partial derivatives of the modified performance index with respect to the unknown matrix elements. Differentiating (23) with respect to Y yields (16). Differentiating (23) with respect to P gives the following adjoint equation:

$$Y = M'YM + U \tag{24}$$

If the symmetric matrices P and Y are partitioned according to

$$P = \left[\begin{array}{c|c} P_1 & P_3 \\ \hline P_3' & P_2 \end{array}\right] \qquad Y = \left[\begin{array}{c|c} Y_1 & Y_3 \\ \hline Y_3' & Y_2 \end{array}\right] \tag{25}$$

then the modified performance index in (23) can be rewritten as an explicit function of the matrices F, G, H, and L. Differentiating the resulting expression, with the aid of the formulas in Appendix A, yields

$$\frac{1}{2}\frac{\partial \bar{J}}{\partial F'} = Y_3'(AP_3 - BHP_2) + Y_2(GCP_3 + FP_2) \tag{26}$$

$$\frac{1}{2}\frac{\partial \bar{J}}{\partial G'} = Y_2 GW + [\, Y_3'(AP_1 - BHP_3') + Y_2(GCP_1 + FP_3')\,]\, C' \tag{27}$$

$$\frac{1}{2}\frac{\partial \bar{J}}{\partial H'} = RHP_2 - B'[\, Y_1'(AP_3 - BHP_2) + Y_3(GCP_3 + FP_2)\,] \tag{28}$$

$$\frac{1}{2}\frac{\partial \bar{J}}{\partial L'} = E(LP_2 - KP_3) \tag{29}$$

Recall that not all the elements of the matrices F, G, H, L are neccessarily design parameters. Hence setting (26)-(29) to zero will most likely result in more equations than unknowns. This difficulty is resolved by applying the following property: Let Fij be a design parameter of the F matrix located in row i and column j. The derivative of the performance index with respect to this design parameter is simply the element of the matrix computed by (26) that is located in row i and column j. In other words

$$\frac{\partial \bar{J}}{\partial F_{ij}} = \left[\frac{\partial \bar{J}}{\partial F'}\right]_{ij} \qquad (30)$$

Equations similar to (30) apply to (27)-(29).

The approach described by (26)-(30) for computing the partial derivatives eliminates the need of differentiating the modified performance index seperately for each design parameter. Hence, this approach is especially useful for computer programming purposes.

The equations obtained by setting (26)-(29) to zero can not be solved analytically for the design parameters in the matrices F, G, H, L. Based on the gradient of the modified performance index, the design parameters are, instead, adjusted iteratively to converge toward the minimum solution. This technique is the steepest descent method and the key steps in the numerical solution are briefly described below.

1. Make an intial guess of F, G, H, L.

2. Solve (16) for P.

3. Solve (24) for Y.

4. Use (25)-(29) to compute the partial derivatives or gradient matrices. Those elements of the 4 gradient matrices that are not the design parameters should be set to zero.

5. Compute a new set of design parameters according to

$$\begin{aligned} F &= F - b(\partial\bar{J}/\partial F') \\ G &= G - b(\partial\bar{J}/\partial G') \\ H &= H - b(\partial\bar{J}/\partial H') \\ L &= L - b(\partial\bar{J}/\partial L') \end{aligned} \qquad b>0 \qquad (31)$$

The scalar b determines the magnitude of the adjustment in the solution. The value of b is selected large enough to provide rapid convergence and yet small enough to insure an accurate solution.

6. Repeat steps 2-6 using the new values of the matrices F, G, H, L until convergence occurs.

CONCLUSIONS

An attractive method for computer-aided control system design is the steady state LQG theory modified to permit reduced order controllers. To apply this approach the designer has the additional task of specifying the order and configuration of the controller. The advantage is that the resulting controller design is simpler and hence better able to meet the demands of real-time control.

One of the limitations of using the reduced order controller, which is addressed by the this paper, is that the controller states can no longer be interpreted as estimates of the plant states. It is shown that the design method can be extended so that the controller will also estimate key variables. This extension does not alter the basic mathematical structure of the equations to be solved for the controller design.

The extension permits the designer to tradeoff estimation accuracy versus control accuracy by varying the weighting matrices in the performace index. The sensitivity of the closed loop performance to increases in the weighting on the estimation accuracy remains to be investigated.

Special formulas are presented for computing the derivative of the trace operator with respect to a matrix. These formulas are useful in developing a general computer program to implement the design method.

The discrete-time control problem is considered in this paper but the method readily extends to the continuous-time case.

The future use of the method for real-time drone control and other applications appears promising.

Appendix A: MATRIX GRADIENT FORMULAS FOR TRACE OPERATOR

Formulas for computing derivatives of scalar functions with respect to a matrix are listed below. The class of scalar functions is the trace of a matrix product. These formulas are useful in the computation of the gradient used by the steepest descent method in finding the solution that minimizes the performance index.

$$H = \partial tr[F(X)G'] / \partial X'$$

$F = A$	$H = 0$
$F = X$	$H = G$
$F = AX$	$H = A'G$
$F = AX'$	$H = G'A$
$F = XB$	$H = GB'$
$F = X'B$	$H = BG'$
$F = AXB$	$H = A'GB'$
$F = AX'B$	$H = BG'A$
$F = X'AX$	$H = AXG' + A'XG$
$F = XAX'$	$H = GXA' + G'XA$
$F = B'X'AXB$	$H = AXBG'B' + A'XBGB'$
$F = BXAX'B'$	$H = B'G'XA' + B'G'BXA$

REFERENCES

1. Levine,W.S., Johnson,T.L., and Athans,M., "Optimal Limited State Variable Feedback Controllers for Linear Systems," IEEE Transactions on Automatic Control, Vol. AC-16, Dec. 1971, pp. 785-793.

2. Sims,C.S. and Melsa,J.L., "A Fixed Configuration Approach to the Stochastic Linear Regulator Problem," Proceedings of the Joint Automatic Control Conference, June 1970, pp. 708-712.

3. Belanger,P.R. and Chung,R., "Design of Low Order Compensators by Gain Matching," IEEE Transactions on Automatic Control, Vol. AC-21, Aug. 1976, pp. 627-629.

4. Martin,G.D. and Bryson Jr.,A. "Attitude Control of a Flexible Spacecraft," AIAA J.Guidance and Control, Vol.3, No.1, Jan.-Feb. 1980, pp. 37-41.

5. Kwakernaak,K. and Sivan,R., Linear Optimal Control Systems, Wiley-Interscience, New York, 1972, pp. 427-436.

6. Denery,D.G., "Identification of System Parameters from Input-Output Data with Applications to Air Vehicles," Thesis, Stanford Univ., May 1971.

HIGH VALUE TARGET MODELING
FOR TACTICAL DECISION AIDS

Richard A. Weiss and
Lewis E. Link
Environmental Laboratory
U. S. Army Engineer Waterways Experimental Station
Vicksburg, MS 39180

ABSTRACT. As part of a program to select an optimal tactical weapon system for specified environmental conditions a procedure is presented for predicting the apparent infrared image of a target in a background that would be seen by an infrared sensor in a missile or airplane. Environmental effects are introduced through the exchange of energy between the target, atmosphere, and background by radiation, conduction, convection, and latent heat. The diurnal variation of the temperature and radiance of target and background surfaces are calculated including the effect of target orientation relative to the sun. Numerical calculations are presented for temperatures and radiances of the facets of a petroleum storage tank.

INTRODUCTION. In any future conflict, technology may provide us with our most effective advantage over a numerically superior foe. Technology in this sense infers not only more capable and sophisticated systems, but also the ability to use them effectively. Emphasis on high technology has generated numerous, remarkably capable systems exemplified by the list of Army electro-optical dependent systems given in Table 1a. These devices and more advanced systems under development will provide significantly enhanced surveillance, target acquisition, and terminal homing capabilities. Doctrine and tactics to exploit the advantages these systems provide must account for their increased sensitivity to the operational environment, the price tag for extended performance and increased autonomy.

The European environment can be especially hostile to electro-optical systems. Table 1b illustrates that visual systems are nearly useless in the winter to detect and lock on to a tank target at a range of 2750 m (1). Infrared (IR) systems fare better; however, they are still ineffective for a significant fraction of the time, especially in the winter. A synopsis of West German weather shows that in the winter overcast cloud conditions occur 70 to 80 percent of the time (Table 1c) (1). Cloud cover can severely impact visible and IR system performance by obscuration (reduction of atmospheric transmission) and for IR systems a suppression of target-background contrast.

Obviously, a single high technology system is not always the optimum choice for a particular mission. It is imperative to know when they will work and when they will not, which systems will be most effective in an anticipated set of conditions, and what performance level can be expected for the system selected. This requires the implementation of the second component of technology mentioned in the first paragraph, the ability to project system performance and optimize the effectiveness of systems by choosing the most appropriate alternative for the anticipated operational environment. This is the essence of the Tactical Decision Aid (TDA) concept (Figure 1a).

In the case of advanced electro-optical controlled weapon systems the concept embraces the development of analytical methods to forecast system performance for the conditions anticipated at the time of target engagement. This includes the type of target, the terrain conditions (background) surrounding the target, weather conditions, and atmospheric transmission. It is in essence an analysis of sensor-environment interactions coupling forecasts of future weather conditions, atmospheric transmission, and the dynamic response characteristics of targets and backgrounds using mathematical algorithms that mimic the most significant interactions. The TDA is a projection of performance for alternative systems, formulated to clearly illustrate which system to use, when to use it, and how well it will work. As such, the TDA is a near-real time mini-war game that strives to optimize effectiveness of our high technology systems.

The prediction of the performance of a sophisticated system must, realistically, account for some rather complex phenomena. For this reason, the basic models should be based on physical principles, and should predict the strength of the intrinsic signal of the target and its attenuation during transmission through the environment to the weapon sensor. The environment affects both the intrinsic target signal and the apparent signature. The prediction schemes must be easily executed and simple to use, and therefore the final package, from the outward appearance, may not reflect the true character. and sensitivities of the internal analytical relationships used to formulate the prediction. For this reason the user must be familiar with the structure of the procedures and the potential limitations of the analytical tools used.

The performance of modern weapon systems depends on the acquisition of target location and appearance data. Weather conditions affect the appearance of a target relative to its background. Prediction of this appearance is crucial to the selection of a weapon system. It is desirable to be able to predict the image of a target in different modes of energy transfer such as radar, infrared, millimeter and submillimeter waves, and others. A model that predicts the characteristic signatures of a target in various forms of energy transfer and in all weather conditions is needed to project system performance. Some weapon systems will be better suited for a particular battlefield environment than others. Modern defense systems are fast and accurate thereby limiting the time available for the on site selection of the appropriate weapon system.

The TDA concept is based on the idea that a particular weapon system can be selected which is best suited for locating and destroying a specified target under expected environmental conditions. Ideally this selection would be made before the launch of the weapon system. The probability of success of a weapon is determined in part by environmental factors. These probabilities are used to assess alternatives and to select an optimum system. The anticipated appearance of the target with respect to the background would be used also to brief a pilot or for insertion into the memory of a missile computer.

The types of environmental information required are: weather, terrain elevations, soil moisture content, vegetation, soil and vegetation temperature, position of the sun in the sky, latitude of target, atmospheric dust, smoke, and smog. The environmental factors can be obtained from a variety of sources.

For instance, weather conditions can be obtained from satellite pictures, terrain elevations, and slopes from photogrammetry, soil moisture, and the temperature of soil and vegetation from infrared photographs, and atmospheric dust, smoke, and smog from measurements of light scattering and attenuation (2). However, systematic methods for obtaining the data to support combat operations remains a problem.

A complete environment-target model would predict the quality and strength of the target and background signatures for different modes of energy transfer and different detectors. These signatures will vary with the time of day as well as with environmental conditions. On the basis of this information an optimal weapon system can be selected and programmed to recognize and attack a target in a background for specified weather conditions and at any time of day or night (Figures 1b and 1c).

OBJECTIVE. The objective of this paper is to present details of a concept for predicting thermal signatures of high-value targets and their backgrounds and to illustrate the role of such a capability as a critical component in a TDA procedure. The TDA concept is first discussed within the perspective of the overall structure, sources of inputs, and the sophistication of critical component algorithms. Through this discussion an effort is made to outline the critical component of TDA procedures and to illustrate the inherent complexity required to forecast the performance of modern weapons systems.

The concepts for high-value target and background thermal signature prediction is presented in detail. High value targets are defined as large critical facilities such as airfields, POL storage, bridges, power plants, and rail yards. While the procedures presented are generic in nature, the high-value target model described is being developed for the Air Force Armaments Laboratory and Air Force Geophysics Laboratory as a component to their TDA program (Figure 1d).

A HIGH VALUE TARGET TDA. A procedure for projecting the performance of an IR weapons system design to detect and lock-on high value targets would require four major components: a target-background signature model, a sensor characteristics model, an atmospheric transmission model, and systems models to tie together the basic component models and provide a means to estimate system performance for the scenario described by the input data.

The target-background signature model function is to project the inherent (at the target) appearance of the target within the background for specified weather conditions. If the weapons system of interest detects and locks-on by a simple contrast criteria the model description of target and background geometry may remain quite simple. The average radiance of the target and the background may be sufficient information to project performance. If background complexity is a factor in sensor performance or the target has specific parts that are significantly hotter than the remainder of the structure of facility a more complex description of the target and background are required. This may require supplementing an overall average contrast value with information on background complexity (clutter level) and/or target hot spots that may be more critical to system performance than the average contrast value. If an imaging sensor is being considered the full geometric complexities of the target and background must be included. The signature simulation then expands essentially to a full scene simulation.

The actual signature calculation consists of an energy budget analysis of the various natural background and target surfaces that would be within the field of view of the sensor. The energy fluxes are driven by the characteristics of the surfaces and the weather. While one dimensional models may adequately handle surfaces for simple targets and contrast dependent sensors, consideration of energy fluxes for scene simulation requires much more complex models that allow energy exchanges between target and background surfaces as dictated by their orientation and proximity.

Sensor characteristic models essentially describe the interaction of sensor components with the received signals from the target and background. The sophistication of the model is in direct proportion to the sophistication of the sensing device and any internal logic used to detect, classify, and lock-on to a target. Once again a contrast seeker is more simply simulated than an imaging system. The first requires only a signal-to-noise analysis while the later requires consideration of the radiance, geometry, texture, and association of the target and background.

Atmospheric transmission models are needed to project the inherent target and background signatures estimated at the location of the target to the signatures received at the position(s) of the sensor system. For modeling realistic battlefield conditions it is necessary to include the effects of natural atmospheric absorption and scattering phenomena as well as the very dramatic impacts of battlefield induced contaminants such as dust and smoke.

The systems performance model integrates the previously described component models to mimic the entire system from signature generation, through atmospheric attenuation, to interaction with the sensor. It acts as an executive program for providing inputs to the specific component models, transferring one component's output to another component for further processing and finally shaping the output of the entire system. While the systems performance model is the glue that ties the components together it is the complexity and capabilities of the component models that really dictate the capabilities of the overall TDA procedure and its limitations.

INPUTS AND THEIR SOURCES. The individual component models require some common and some unique inputs. The sensor characteristic model primarily requires indicies related to sensor performance such as electrical bandwidth, spatial resolution, spectral sensitivity, signal processing transforms, and logic/ classified criteria.

The atmospheric transmission model requires a description of atmospheric aerosol and molecular constituents. Inputs should be tied closely to regional atmospheric characteristics, local weather conditions, and battlefield activities. This can be handled most easily by using a series of preestablished look-up tables representing a range of anticipated atmospheric conditions. In this manner, an appropriate atmospheric condition is selected from a menu and the system performance code provides the required information to the atmospheric model to make representative atmospheric transmission calculations.

The major input requirement is for the target-background signature model. Calculation of energy fluxes requires a description of the thermal and physical characteristics of the target and background surfaces of most significance. These are basically static parameter values that can be provided from look-up tables keyed to specific surface types and conditions. Some surface thermal properties, such as the thermal conductivity of soil, are dynamic and correct parameter value estimates require a determination of the state of the surface for the time of signature forecast and in many instances some prior period.

Weather is the principal driver of the energy fluxes as well as the state of the surfaces. As such it becomes the most critical input to the TDA procedure. Wind speed, air temperature, humidity, cloud cover, rainfall, and solar insolation data on a timely basis are critical to accurate calculation of surface IR signatures. Since the object is to estimate signatures and sensor performance at some time in the future, the weather data input by necessity must result from forecasts or extensions of measured data. As the time lapse between the last measured weather input and the time of engagement increases, the accuracy of the forecast weather data input to the model clearly diminishes. The TDA procedures must be designed to provide any advantage possible to offset the possible inaccuracies in the weather inputs.

Since weather data are critical to the execution of a high-value target TDA a near-real time weather data acquisition, reduction, and processing capability is needed to provide the procedure with the best possible information. Progress in mesoscale weather models would add materially to the reliability of such TDA performance projections.

INFRARED TARGET AND ENVIRONMENT MODEL FOR TDA. A specific example of a target-environment model that can be used in a TDA program is the IR imaging of high value targets such as airfield runways and industrial complexes such as petroleum storage tanks. The predicted image of a target in a background should agree with the image displayed by an infrared detector in a missile or airplane. This prediction tells the pilot or missile computer what type of actual IR image to expect during a tactical military mission. This will reduce the recognition time for a target which may be camouflaged and not easily seen in visible light.

The IR imaging model consists of two basic components: 1) a target and background IR radiance model, and 2) a scene generation model which projects the intrinsic target and background radiances into the focal plane of the sensor of the attacking missile or airplane. The target and background radiances are calculated using the computer program Terrain Surface Temperature Model (TSTM), and the infrared scenes are generated from a computer program called Scene Generator (SCNGEN) (3,4). The SCNGEN computer program requires as input the target and background radiances generated by TSTM. Both component models require input data from the environment (Figures 2a and 2b).

BASIC PHOTOMETRIC DEFINITIONS. The spectral radiance of a target surface is defined as the amount of power radiated per unit area, per unit solid angle, and per unit wavelength as follows (5-7)

$$dP = N_\lambda \cos \theta \, dA \, d\Omega \, d\lambda \qquad (1)$$

where

P = emitted power, watts

N_λ = spectral radiance, watts $\cdot$ m^{-2} $\cdot$ µm^{-1} $\cdot$ sr^{-1}

θ = angle between direction of observation and the normal to the target surface, radians

A = area of target surface, m^2

Ω = solid angle, steradians (sr)

λ = wavelength, µm

The radiance of a surface is defined as

$$N = \int_{\lambda_1}^{\lambda_2} N_\lambda \, d\lambda, \text{ watts} \cdot \text{m}^{-2} \cdot \text{sr}^{-1} \qquad (2)$$

and is associated with specified wavelength intervals. The target facet radiances used in TSTM are associated with the 3-5 and 7-14 micron wavebands.

The magnitude of the IR radiance associated with the background and target is due to a number of physical processes associated with the absorption, emission, and reflection of electromagnetic energy. The total IR radiance associated with target and background surfaces consists of a thermal component and a reflected component. The thermal radiance depends on the absolute temperature of the target and background facets.

GROUND TEMPERATURE. The surface temperature of the background is calculated by the TSTM (3). Several physical processes are included in this model (Figure 3a):

a. radiative absorption and thermal emission of the surface

b. conduction of heat in target and background

c. conduction and convection by the atmosphere

d. latent heat of evaporation

The equilibrium surface temperature is determined by satisfying an energy flow budget equation at the surface of a target or background (3,8). The energy budget includes the absorption of a SW insolation component and a LW atmospheric term. The ground temperature is calculated from the following two major irradiance terms:

a. Direct SW Insolation on Ground

The SW solar irradiance on the ground is given by (3)

$$H_o^g(SW) = S_o W_1 \cos Z_s \tag{3}$$

where

$H_o^g(SW)$ = solar irradiance on ground, watts · m^{-2}

S_o = solar constant = 1395.0 watts · m^{-2}

W_1 = cloud cover function (reference 3)

Z_s = solar zenith angle

b. Direct LW Atmospheric Irradiance on Ground

The atmospheric LW irradiance on the ground is given by the Brunt equation (3)

$$H_{atm}^g(LW) = W_2(T_a, e_a) \tag{4}$$

where

H_{atm}^g = LW irradiance on ground, watts · m^{-2}

$W_2(T_a, e_a)$ = air temperature and water vapor function (reference 3)

T_a = air temperature

e_a = water vapor pressure

The LW irradiance is diffuse. The $W_2(T_a, e_a)$ function contains empirical coefficients.

IRRADIANCE ON TARGET FACET. The TSTM was originally designed to predict the ground temperature and the temperatures of horizontal surfaces. This model has been modified to describe vertical surfaces by including the important radiative interchanges between the vertical target, the atmosphere, and the background.

The radiative transfer at the target surfaces has longwave (LW) and shortwave (SW) components and include the following five major irradiance components (Figures 3b-3d):

a. Direct SW Insolation on Target

The direct SW solar irradiance on a target surface is given by

$$H_o^{tar}(SW) = S_o W_1 \cos \phi_t \tag{5}$$

where

$H_o^{tar}(SW)$ = solar irradiance on target surface, watts · m^{-2}

ϕ_t = angle between surface normal and the direction of the sun

A simple analysis shows that

$$\cos \phi_t = \sin Z_s \sin (SL) \cos (S_{AZ} - SL_{AZ}) + \cos Z_s \cos (SL) \qquad (6)$$

where

SL = slope angle of target surface

S_{AZ} = solar azimuth angle

SL_{AZ} = azimuth angle of normal to target surface

b. SW Solar Radiation Reflected from Background to Target

The SW irradiance on the target surface due to reflection from the ground is given by

$$H_{ref,g}^{tar}(SW) = r_g(SW)S_oW_1 \cos Z_s \cos \phi_t^R \qquad (7)$$

where

$r_g(SW)$ = SW reflected coefficient of ground = 1 - $\varepsilon_g(SW)$

$\varepsilon_g(SW)$ = SW absorptivity of ground

ϕ_t^R = angle between surface normal and the reflected sun ray

The SW reflectivity may depend on the solar zenith angle, especially for the case of water surfaces. Simple geometry shows that

$$\cos \phi_t^R = \sin Z_s \sin (SL) \cos (S_{AZ} - SL_{AZ}) - \cos Z_s \cos (SL) \qquad (8)$$

c. Direct LW Atmospheric Irradiance on Target Surface

The atmospheric LW irradiance on a target surface is assumed to be given by the Brunt equation as in the case of the LW irradiance on the ground (equation 4)

$$H_{atm}^{tar}(LW) = W_2(T_a,e_a) \qquad (9)$$

where $H^{tar}_{atm}(LW)$ = direct LW irradiance on target surface, watts · m^{-2}

d. LW Atmospheric Radiation Reflected from Ground to Target

Because the atmospheric LW radiation is diffuse in nature the radiation reflected from the ground will also be diffuse, and the irradiance on a raised target surface is given by

$$H^{tar}_{ref,g}(LW) = \frac{1}{A_t} \iint \frac{N^g_{ref}(LW) \cos \theta_t \cos \theta_g}{r^2} \, dA_g \, dA_t \qquad (10)$$

where the diffuse reflected LW ground radiance is given by

$$N^g_{ref}(LW) = \frac{r_g(LW)}{\pi} H^g_{atm}(LW) \qquad (11)$$

and where

A_t = target area

θ_t = angle between normal to target and reflected ray

θ_g = angle between normal to ground and incident ray

r = variable distance between points on ground and on target surface

A_g = area of ground

$r_g(LW)$ = LW reflectivity of ground = $1 - \varepsilon_g(LW)$

$\varepsilon_g(LW)$ = LW absorptivity of ground = LW emissivity of ground

For the case of a homogeneous half-plane background the integral in equation 10 simplifies to the result

$$H^{tar}_{ref,g}(LW) = \frac{\pi}{2} N^g_{ref}(LW) \qquad (10a)$$

e. LW Irradiance on Target Due to Thermal Radiance of the Ground

The diffuse LW irradiance on a raised target due to the thermal radiance of the ground is given by

$$H^{tar}_{ther,g}(LW) = \frac{1}{A_t} \iint \frac{N^g_{ther}(LW) \cos \theta_t \cos \theta_g}{r^2} \, dA_g \, dA_t \qquad (12)$$

where the thermal ground radiance is obtained from Planck's law as

$$N_{ther}^{g}(LW) = \varepsilon_{g}(LW) C_{1}^{N} \int_{\lambda_1}^{\lambda_2} \frac{d\lambda}{\lambda^{5}\left[\exp\,(C_{2}^{N}/\lambda T) - 1\right]} \tag{13}$$

where

$N_{ther}^{g}(LW)$ = thermal radiance of ground

$C_{1}^{N} = 1.192 \times 10^{8}$ watts · $(\mu m)^{4} \cdot m^{-2} \cdot sr^{-1}$

$C_{2}^{N} = 1.4338 \times 10^{4}\ \mu m \cdot {}^{o}K$

T = ground temperature, °K

The IR wavebands are taken to be 3-5 and 7-14 microns. For the case of a homogeneous half-plane background the integral in equation 12 simplifies to

$$H_{ther,g}^{tar}(LW) = \frac{\pi}{2} N_{ther}^{g}(LW) \tag{12a}$$

The total SW irradiance on the target surface is obtained from equations 5 and 7 to be

$$H^{tar}(SW) = H_{o}^{tar}(SW) + H_{ref,g}^{tar}(SW) \tag{14}$$

The total LW irradiance on the target surface is obtained from equations 9, 10, and 12 to be

$$H^{tar}(LW) = H_{atm}^{tar}(LW) + H_{ref,g}^{tar}(LW) + H_{ther,g}^{tar}(LW) \tag{15}$$

Finally the total irradiance on a target surface is given by

$$H^{tar} = H^{tar}(SW) + H^{tar}(LW) \tag{16}$$

which has a SW and a LW contribution. For horizontal surfaces such as the ground, the target irradiances given in equations 14 and 15 reduce to those given in equations 3 and 4, respectively.

Less important radiative transfer effects are:

a. SW solar radiation reflected from target surface to the background and reradiated as IR thermal emittance which irradiates target.

b. Atmospheric IR radiation reflected from target surface to the background and reradiated as a thermal IR emittance which in turn irradiates the target.

TARGET FACET TEMPERATURE. The absorbed part of the total irradiance on a target surface enters the calculation of the target surface temperature. The absorbed power per unit is obtained from equations 14 and 15 as

$$H^{tar}_{abs} = \varepsilon_t(SW)H^{tar}(SW) + \varepsilon_t(LW)H^{tar}(LW) \tag{17}$$

where

H^{tar}_{abs} = power absorbed per unit area, watts · m^{-2}

$\varepsilon_t(SW)$ = SW absorptivity of target surface

$\varepsilon_t(LW)$ = LW absorptivity of target surface

It is this total absorbed power that enters the calculation of target facet temperature using the computer program TSTM. The procedure for calculating the target surface temperature using equation 17 is essentially the same as used for calculating the ground temperature.

For surfaces that are not horizontal the orientation of the surface with respect to the direct and reflected radiation components enters the irradiance calculation. This introduces a dependence of the predicted target surface temperature on the slope and azimuth angles of the surface, and a dependence on the solar zenith and solar azimuth angles.

TARGET FACET RADIANCE. The total IR emittance from a target or background surface is the sum of a reflected IR component and a thermal IR emittance which is calculated from the predicted target surface temperature.

a. Thermal IR Radiance of Target

The thermal radiance of a target or background surface is calculated from Planck's black body radiation law as follows (Figure 3d)

$$N^{tar}_{ther}(LW) = \varepsilon_t(LW)C^N_1 \int_{\lambda_1}^{\lambda_2} \frac{d\lambda}{\lambda^5 \left[\exp\ (C^N_2/\lambda T) - 1\right]} \tag{18}$$

where

N^{tar}_{ther} = thermal radiance of target, watts · m^{-2} · sr^{-1}

T = target surface temperature, °K

The IR transmission windows are taken to be 3-5 and 7-14 microns. The important parameters for the determination of thermal radiance are the target surface temperature, the emissivity of the surface, and the transmission wavebands.

b. Reflected IR Radiance

In addition to the thermal IR radiance of a surface there is also a reflected IR radiance which for a vertical surface is due in part to the IR atmospheric radiation reflected directly from the target, and in part to the IR atmospheric radiation reflected from the background and then reflected again from the target surface. There is also an IR reflected component due to the thermal emittance of the background which is reflected by the target surface. The reflected IR radiance is calculated, under the assumption that it is diffuse, in the following manner from equation 15 (Figure 10)

$$N_{ref}^{tar}(LW) = \frac{r_t(LW)}{\pi} H^{tar}(LW) \qquad (19)$$

where $r_t(LW)$ = LW reflectivity = $1 - \varepsilon_t(LW)$.

The total target facet IR radiance is obtained finally as the sum of the thermal radiance and the LW reflected radiance as follows

$$N^{tar}(LW) = N_{ther}^{tar}(LW) + N_{ref}^{tar}(LW) \qquad (20)$$

This is the intrinsic target radiance that serves as input for the determination of the apparent radiance seen by the IR detector.

APPARENT TARGET AND BACKGROUND RADIANCE. The apparent target and background IR radiances measured in the focal plane of an IR detector in an aircraft is less in magnitude than the intrinsic radiances for the following reasons:

a. geometrical attenuation according to the inverse square of the distance between target and detector

b. atmospheric absorption and scattering

c. obscuration due to smoke, clouds, dust, etc.

d. orientation of the missile relative to the ground plan

The synthetic IR imaging of a target in a background will be accomplished by the SCNGEN computer program (4).

The terrain background is represented by several textures such as trees, grass, roads, etc. In order to construct a scene the target must be located within the background, and the detector position, viewing direction, and orientation must be specified. The angular field of view of the detector determines the footprint on the ground plane which is centered about the pierce point of the line-of-sight between sensor and ground plane.

The target surface is approximated by planar triangular facets. Each facet is assigned a value of intrinsic radiance by the TSTM computer program. As discussed earlier in this paper the intrinsic radiance values depend on the orientation of the facets relative to the sun. Facets not visible from the sensor's location and orientation are systematically eliminated by SCNGEN. If more than one target is considered, additional targets are implanted so that occlusion of one target by another is accomplished. The effects of transmission of the IR image through the atmosphere are handled by a version of the LOWTRAN 4 computer program.

NUMERICAL RESULTS. Numerical calculations have been done for the surface temperatures and intrinsic radiances of the facets of a petroleum storage tank (Figure 4a). A petroleum storage tank is constructed of steel plates whose surfaces are painted or rusted to some degree. There is sometimes an airspace over the petroleum, and heat conduction in both petroleum and air was considered. Table 4 gives the thermal conductivities, diffusivities, and specific heats of air, petroleum, and soil, and also gives the LW emissivities and SW absorptivities of the surface coatings considered for this numerical study (3,9).

Figure 4b shows a typical diurnal variation of soil temperature and air temperature. Figure 4c shows horizontal facet temperatures for the top of a petroleum storage tank, and Figures 4d and 4e give the diurnal variation of vertical facet temperature including the dependence on the azimuth angle of the facets. Figure 5a shows a typical diurnal variation of the radiance of a soil background, while Figure 5b gives the radiance of a horizontal facet of a tank. The diurnal variation of the vertical facet radiances and their dependence on azimuth are given in Figures 5c and 5d. Figures 5e through 6b give examples of the expected intrinsic radiance contrast for the facets of a petroleum storage tank.

Figures 4e through 6b show that the predicted surface temperatures adjacent to the airspace and to the petroleum are nearly the same. However, the difference in the values of the thermal inertia (density times specific heat) of air and petroleum suggests that a large temperature difference and phase lag should exist between the air and petroleum temperatures. Preliminary experimental results suggest that this is the case. The TSTM computer program is totally insensitive to the thermal diffusivity (thermal inertia) of the working material (3). This represents an unphysical prediction of the TSTM computer program and is the reason for the nearly identical thermal response of air and petroleum presented in this paper.

The calculation of apparent target-background radiance, as would be seen by an aircraft IR sensor, is in progress using the SCNGEN computer program. When completed an experimental validation program will be carried out by comparing the predicted apparent radiances (IR scene) of a petroleum storage tank with the digitized experimental values of the apparent radiances as given in the Target and Background Information Library System (TABILS) database (Figure 6c).

CONCLUSION. An environmental model and an IR image formation model have been developed which are based on the use of physical principles to describe the basic processes involved in the exchange of energy between the target and the environment. Changing weather conditions affect the energy exchange and alter the appearance of a target in a background. The degree of contrast of the IR image of a target in a background will vary with the hour of day, azimuthal angle of attack, target surface conditions, and local weather conditions. This degree of contrast can be used to predict the performance of alternate IR sensor weapon systems against a specified target and background (Figure 6d).

REFERENCES

1. Kays, M. D., Seagraves, M. A., Monahan, H. H., and Sutherland, R. A., "Qualitative Descriptions of Obscuration Factors in Central Europe," ASL Monograph No. 4, U. S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, N. Mex., Sep 1980.

2. Lintz, Jr., J. and Simonett, D. S., Editors, Remote Sensing of Environment, Addison-Wesley, Reading, Mass., 1976.

3. Balick, L. K., Link, L. E., Scoggins, R. K. and Solomon, J. L., "Thermal Modeling of Terrain Surface Elements," Technical Report EL-81-2, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss., 1981.

4. Baird, A. M., D'Argenio, C. S., Greenwood, F. C., and Bloomer, R., "Computer Generation of Realistic Infrared Scenes of Ground Targets in Environmental Backgrounds," preprint, Analytics Corp., Willow Grove, Pa., 1982.

5. Kruse, P. W., McGlauchlin, L. D., and McQuistan, R. B., Elements of Infrared Technology, John Wiley, New York, 1962.

6. Jamieson, J. A., McFee, R. H., Plass, G. N., Grube, R. H., and Richards, R. G., Infrared Physics and Engineering, McGraw-Hill, New York, 1963.

7. Holter, M. R., Nudelman, S., Suits, G. H., Wolfe, W. L., Zissis, G. J., Fundamentals of Infrared Technology, Macmillan, New York, 1962.

8. Budyko, M. I., Climate and Life, Academic Press, New York, 1974.

9. Heat Transfer and Fluid Flow Data Books, General Electric Co., P. O. Box 43, Schenectady, N. Y., 1975.

TABLE 1

Army Electro-Optical Systems

ARMY E-O DEPENDENT SYSTEMS (OPTICAL)

SYSTEM	WAVEBAND, µm
NIGHT VISION GOGGLES	0.4-0.9
INDIVIDUAL SERVED WEAPON SIGHT	0.4-0.9
M35/E1, M36/E1 PERISCOPES	0.4-0.9
CREW SERVED WEAPON SIGHT	0.4-0.9
HUMAN EYE	0.4-0.7
7 × 50 BINOCULARS	0.4-0.7
DRAGON DAY SIGHT	0.4-0.7
TOW DAY SIGHT	0.4-0.7
GLLD DAY SIGHT	0.4-0.7
TADS DAY SIGHT	0.4-0.7
M-1 DAY SIGHT	0.4-0.7
TTS DAY SIGHT	0.4-0.7

ARMY E-O DEPENDENT SYSTEMS (INFRARED)

SYSTEM	WAVEBAND, µm
TOW NIGHT SIGHT	7.5-12
GLLD NIGHT SIGHT	7.5-12
NODLR NIGHT SIGHT	7.5-12
DRAGON NIGHT SIGHT	7.5-12
TANK THERMAL SIGHT	7.5-12
THERMAL IMAGING SUBSYSTEM	7.5-12
NIGHT CHAPARRAL FLIR SUBSYSTEM	7.5-12
TARGET ACQUISITION DESIGNATION SIGHT	7.5-12
HANDHELD THERMAL VIEWER	3.8-5.5

a.

PERFORMANCE OF VISUAL AND INFRARED DETECTORS - WEST GERMANY

TANK TARGET AT 2750-M RANGE
AND CLOUD CEILING 600-M

DETECTOR	FREQUENCY OF FAILURE JAN	FREQUENCY OF FAILURE JUL
VISUAL	0.86	0.64
INFRARED	0.41	0.13
MILLIMETRE	?	?

b.

WEST GERMANY CLOUD COVER

CLOUD COVER	PERCENT FREQUENCY SUMMER	PERCENT FREQUENCY WINTER
CLEAR TO SCATTERED 0-30%	30	10
BROKEN TO OVERCAST 70-100%	50	70-80

c.

Thermodynamic and Radiometric
Properties of Materials

Thermodynamic Properties*

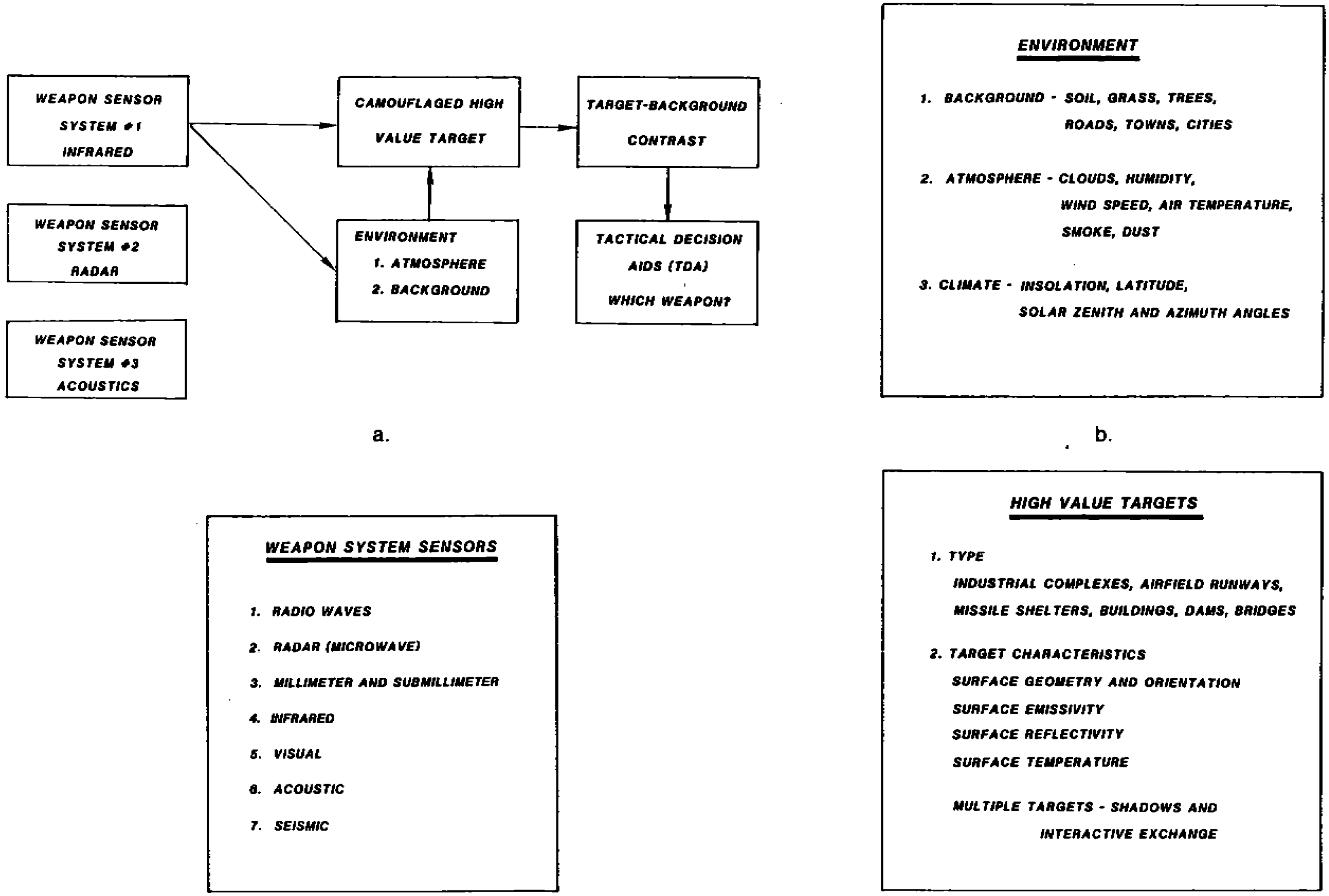
Material	Thermal Conductivity (cal/cm·min·°K)	Thermal Diffusivity (cm^2/min)	Specific Heat, C_p (cal/gm·°K
Soil	0.20	0.3	0.3
Asphalt Pavement	0.106	0.28	0.3
Water	0.084	0.084	1.0
Petroleum	0.02	0.033	0.6
Air	0.0036	12.9	0.243

Radiometric Properties*

Material	Shortwave Absorptivity	Longwave Emissivity
Rust	0.94	0.90
White paint	0.20	0.91
Aluminum paint	0.20	0.40
Soil	0.90	0.70
Asphalt pavement	0.90	0.96

* These properties may vary considerably with composition and weather conditions.

d.


WEAPON SENSOR SYSTEM #1 INFRARED
WEAPON SENSOR SYSTEM #2 RADAR
WEAPON SENSOR SYSTEM #3 ACOUSTICS
CAMOUFLAGED HIGH VALUE TARGET
ENVIRONMENT
1. ATMOSPHERE
2. BACKGROUND
TARGET-BACKGROUND CONTRAST
TACTICAL DECISION AIDS (TDA)
WHICH WEAPON?

a.

ENVIRONMENT

1. BACKGROUND - SOIL, GRASS, TREES,
ROADS, TOWNS, CITIES

2. ATMOSPHERE - CLOUDS, HUMIDITY,
WIND SPEED, AIR TEMPERATURE,
SMOKE, DUST

3. CLIMATE - INSOLATION, LATITUDE,
SOLAR ZENITH AND AZIMUTH ANGLES

b.

WEAPON SYSTEM SENSORS

1. RADIO WAVES
2. RADAR (MICROWAVE)
3. MILLIMETER AND SUBMILLIMETER
4. INFRARED
5. VISUAL
6. ACOUSTIC
7. SEISMIC

c.

HIGH VALUE TARGETS

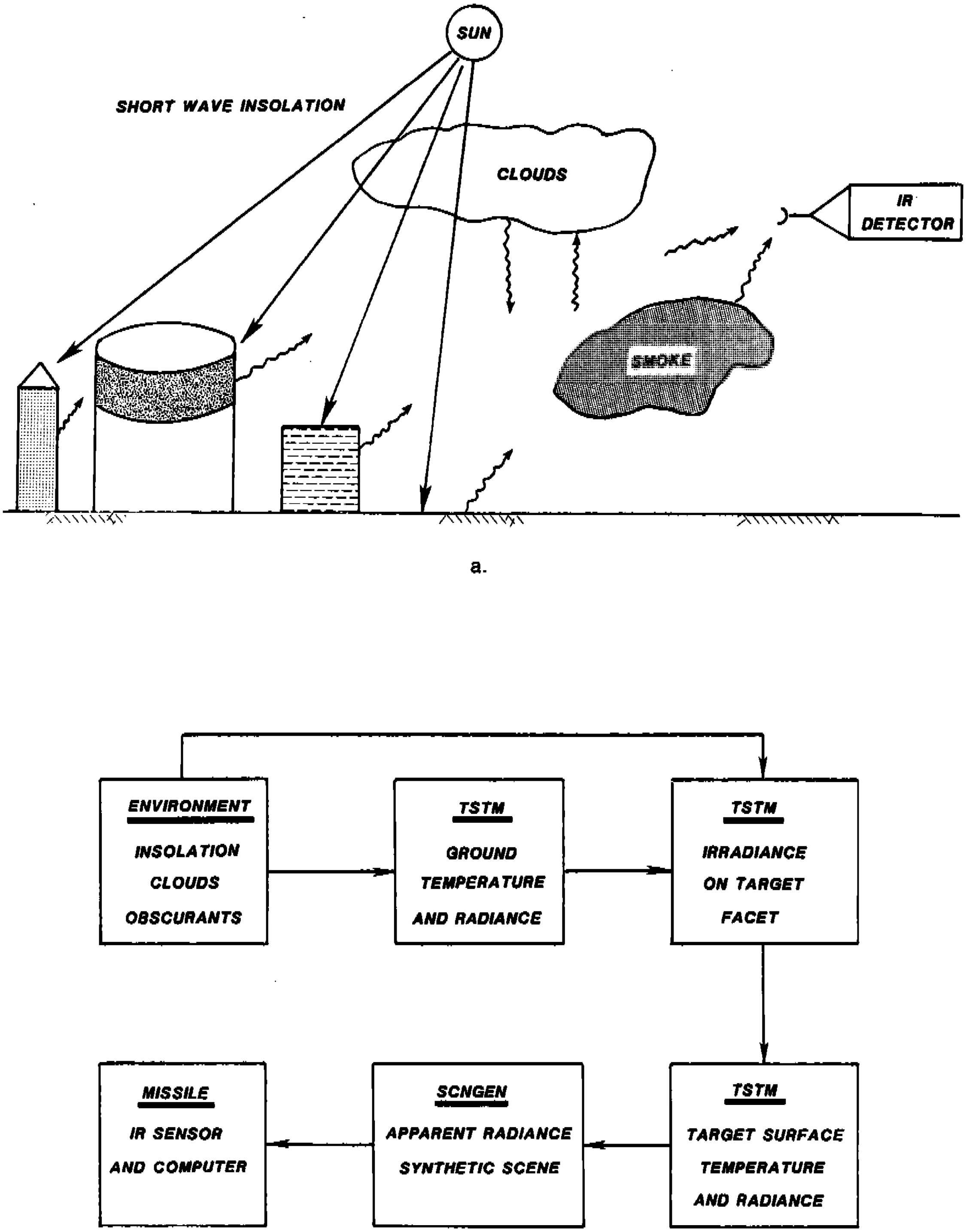
1. TYPE
INDUSTRIAL COMPLEXES, AIRFIELD RUNWAYS,
MISSILE SHELTERS, BUILDINGS, DAMS, BRIDGES

2. TARGET CHARACTERISTICS
SURFACE GEOMETRY AND ORIENTATION
SURFACE EMISSIVITY
SURFACE REFLECTIVITY
SURFACE TEMPERATURE

MULTIPLE TARGETS - SHADOWS AND
INTERACTIVE EXCHANGE

d.

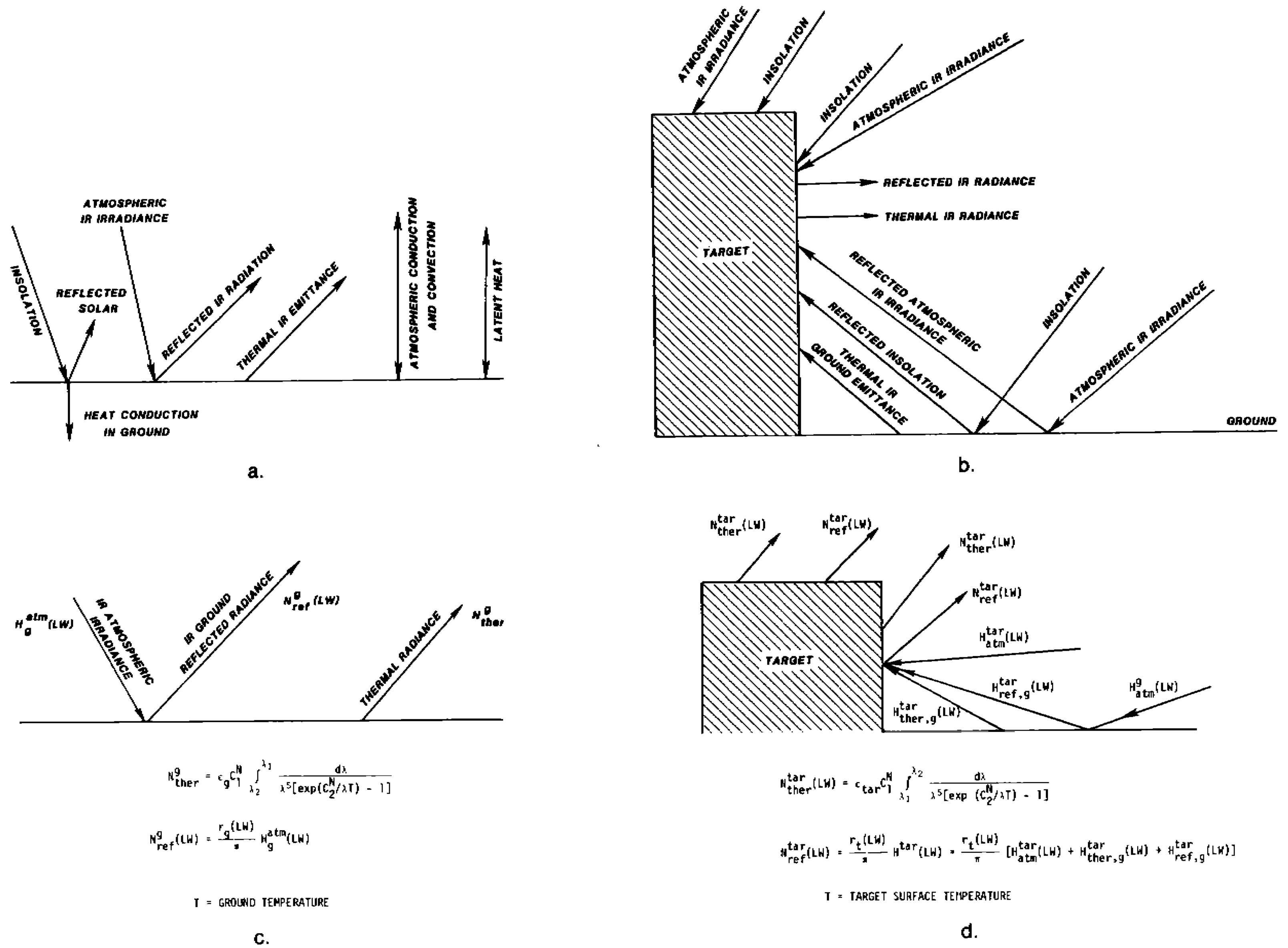
FIGURE 1


SUN
SHORT WAVE INSOLATION
CLOUDS
IR DETECTOR
SMOKE
ENVIRONMENT
INSOLATION
CLOUDS
OBSCURANTS
TSTM
GROUND
TEMPERATURE
AND RADIANCE
TSTM
IRRADIANCE
ON TARGET
FACET
MISSILE
IR SENSOR
AND COMPUTER
SCNGEN
APPARENT RADIANCE
SYNTHETIC SCENE
TSTM
TARGET SURFACE
TEMPERATURE
AND RADIANCE

a.

b.

FIGURE 2



$$N_{ther}^{g} = \epsilon_g C_1^N \int_{\lambda_2}^{\lambda_1} \frac{d\lambda}{\lambda^5[\exp(C_2^N/\lambda T) - 1]}$$

$$N_{ref}^{g}(LW) = \frac{r_g(LW)}{\pi} H_g^{atm}(LW)$$

T = GROUND TEMPERATURE

c.

$$N_{ther}^{tar}(LW) = \epsilon_{tar} C_1^N \int_{\lambda_1}^{\lambda_2} \frac{d\lambda}{\lambda^5[\exp\,(C_2^N/\lambda T) - 1]}$$

$$N_{ref}^{tar}(LW) = \frac{r_t(LW)}{\pi} H^{tar}(LW) = \frac{r_t(LW)}{\pi} [H_{atm}^{tar}(LW) + H_{ther,g}^{tar}(LW) + H_{ref,g}^{tar}(LW)]$$

T = TARGET SURFACE TEMPERATURE

d.

FIGURE 3

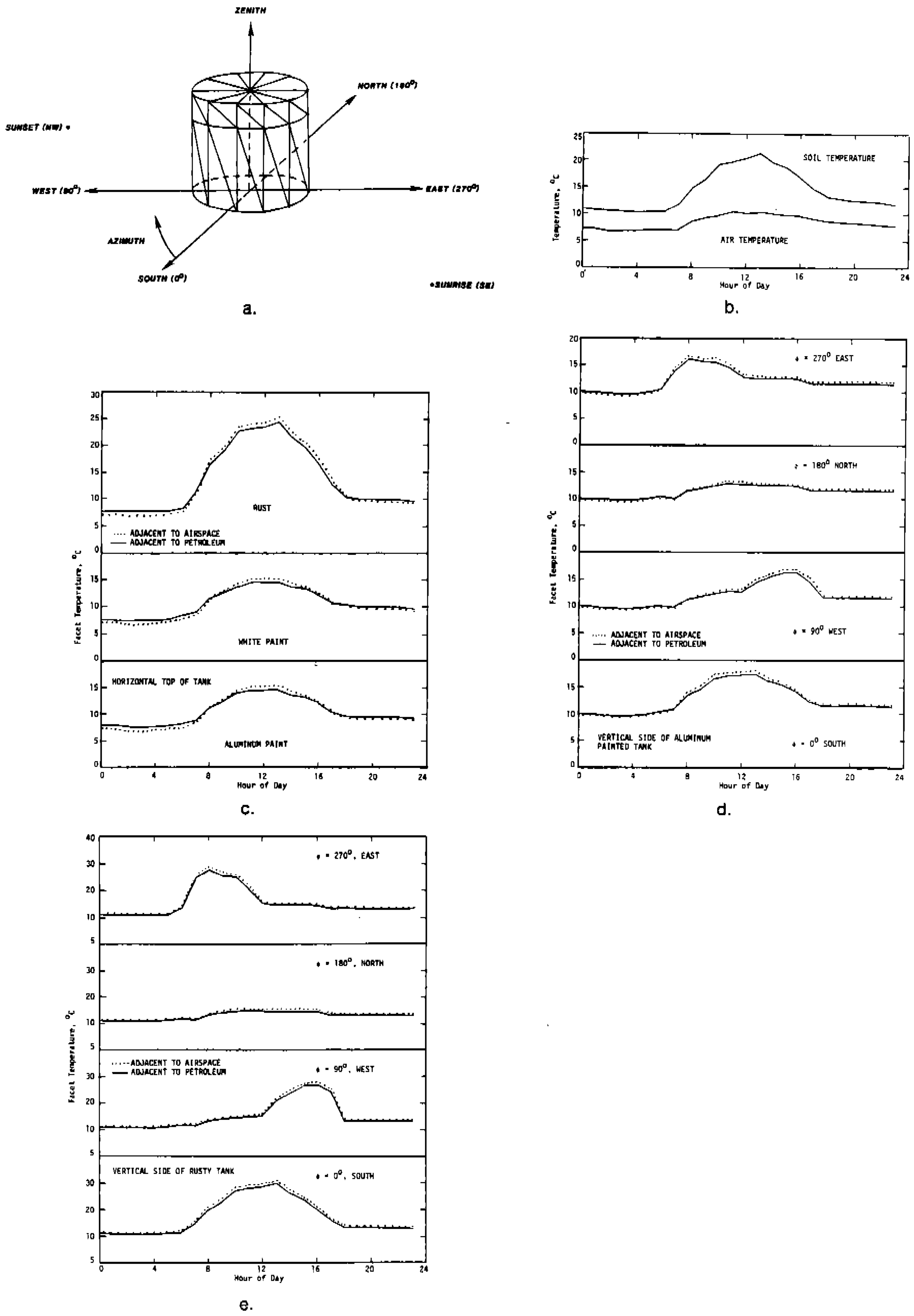

ZENITH
NORTH (180°)
SUNSET (NW)
WEST (90°)
EAST (270°)
AZIMUTH
SOUTH (0°)
SUNRISE (SE)
a.
SOIL TEMPERATURE
AIR TEMPERATURE
Temperature, °C
Hour of Day
b.
RUST
ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
WHITE PAINT
HORIZONTAL TOP OF TANK
ALUMINUM PAINT
Facet Temperature, °C
Hour of Day
c.
φ = 270° EAST
φ = 180° NORTH
ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
φ = 90° WEST
VERTICAL SIDE OF ALUMINUM PAINTED TANK
φ = 0° SOUTH
Facet Temperature, °C
Hour of Day
d.
φ = 270°, EAST
φ = 180°, NORTH
ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
φ = 90°, WEST
VERTICAL SIDE OF RUSTY TANK
φ = 0°, SOUTH
Facet Temperature, °C
Hour of Day
e.

FIGURE 4

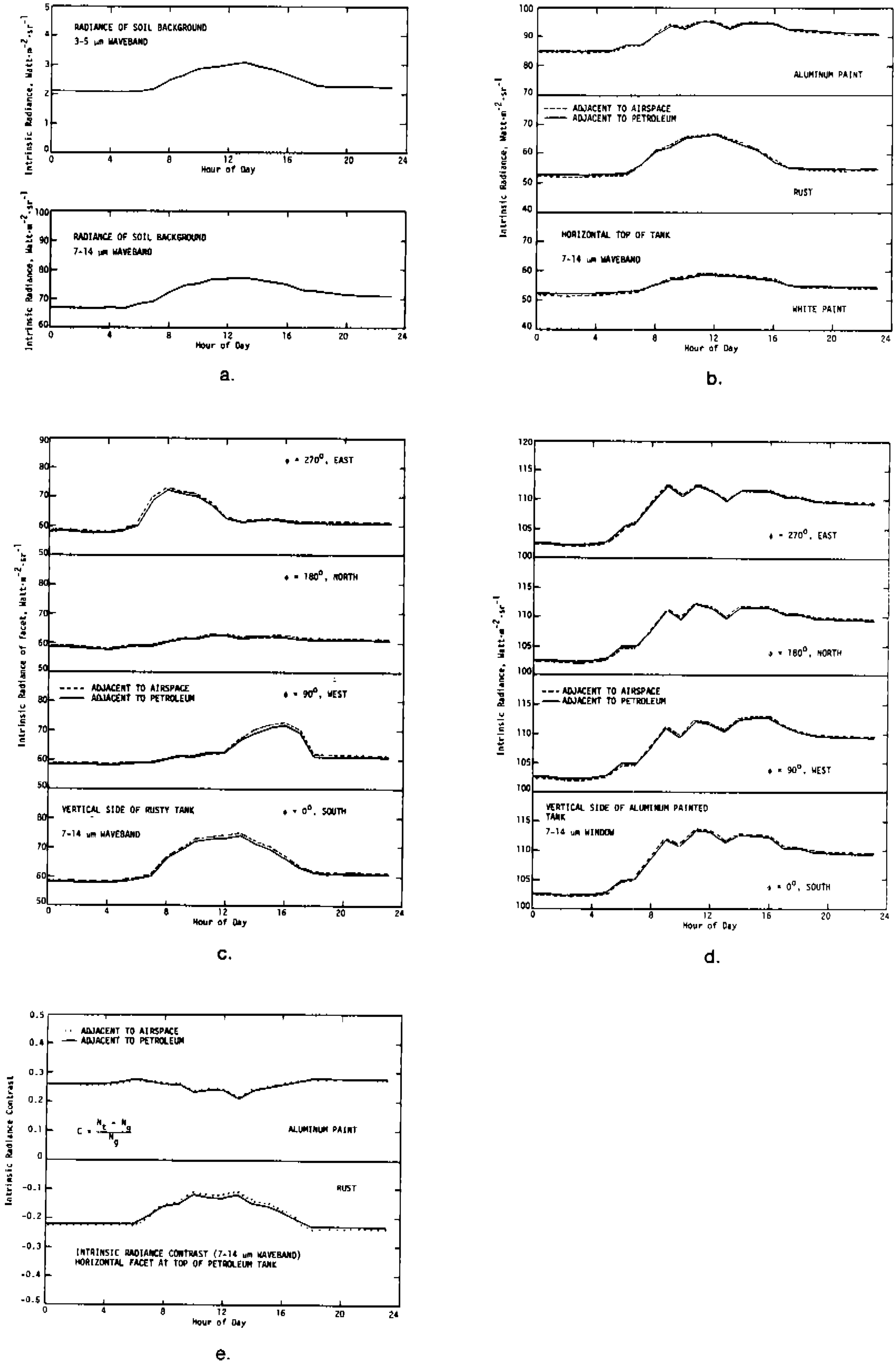
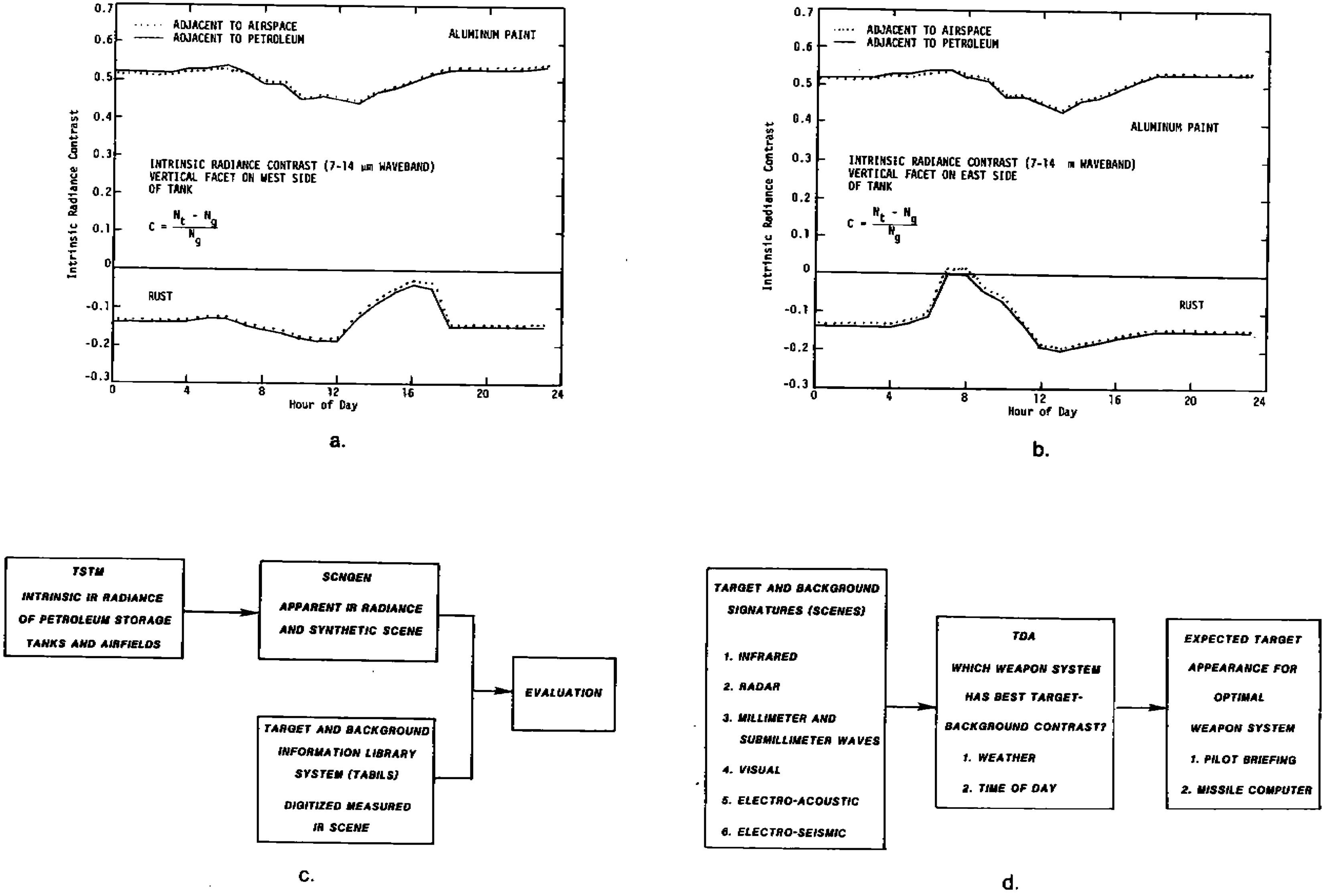

RADIANCE OF SOIL BACKGROUND
3-5 μm WAVEBAND
Intrinsic Radiance, Watt·m^{-2}·sr^{-1}
Hour of Day
RADIANCE OF SOIL BACKGROUND
7-14 μm WAVEBAND
a.
ALUMINUM PAINT
ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
RUST
HORIZONTAL TOP OF TANK
7-14 μm WAVEBAND
WHITE PAINT
b.
φ = 270°, EAST
φ = 180°, NORTH
φ = 90°, WEST
φ = 0°, SOUTH
VERTICAL SIDE OF RUSTY TANK
7-14 μm WAVEBAND
Intrinsic Radiance of Facet, Watt·m^{-2}·sr^{-1}
c.
VERTICAL SIDE OF ALUMINUM PAINTED TANK
7-14 μm WINDOW
d.
Intrinsic Radiance Contrast
$C = \frac{N_t - N_g}{N_g}$
ALUMINUM PAINT
RUST
INTRINSIC RADIANCE CONTRAST (7-14 μm WAVEBAND)
HORIZONTAL FACET AT TOP OF PETROLEUM TANK
e.

FIGURE 5



ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
ALUMINUM PAINT
INTRINSIC RADIANCE CONTRAST (7-14 μm WAVEBAND)
VERTICAL FACET ON WEST SIDE
OF TANK
C = (N_t - N_g) / N_g
RUST
Intrinsic Radiance Contrast
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
0
4
8
12
16
20
24
Hour of Day
a.
ADJACENT TO AIRSPACE
ADJACENT TO PETROLEUM
ALUMINUM PAINT
INTRINSIC RADIANCE CONTRAST (7-14 m WAVEBAND)
VERTICAL FACET ON EAST SIDE
OF TANK
C = (N_t - N_g) / N_g
RUST
Intrinsic Radiance Contrast
Hour of Day
b.
TSTM
INTRINSIC IR RADIANCE
OF PETROLEUM STORAGE
TANKS AND AIRFIELDS
SCNGEN
APPARENT IR RADIANCE
AND SYNTHETIC SCENE
EVALUATION
TARGET AND BACKGROUND
INFORMATION LIBRARY
SYSTEM (TABILS)
DIGITIZED MEASURED
IR SCENE
c.
TARGET AND BACKGROUND
SIGNATURES (SCENES)
1. INFRARED
2. RADAR
3. MILLIMETER AND
SUBMILLIMETER WAVES
4. VISUAL
5. ELECTRO-ACOUSTIC
6. ELECTRO-SEISMIC
TDA
WHICH WEAPON SYSTEM
HAS BEST TARGET-
BACKGROUND CONTRAST?
1. WEATHER
2. TIME OF DAY
EXPECTED TARGET
APPEARANCE FOR
OPTIMAL
WEAPON SYSTEM
1. PILOT BRIEFING
2. MISSILE COMPUTER
d.

FIGURE 6

ORTHOGONAL SCHEMES FOR STRUCTURAL OPTIMIZATION *

M. W. BERRY,‡ M. T. HEATH,† R. J. PLEMMONS‡ and R. C. WARD†

Abstract. Historically there are two principal methods of matrix structural analysis, the displacement (or stiffness) method and the force (or flexibility) method. In recent times the force method has been used relatively little because the displacement method has been deemed easier to implement on digital computers, especially for large sparse systems. The force method has theoretical advantages, however, for multiple redesign problems or nonlinear elastic analysis because it allows the solution of modified problems without restarting the computation from the beginning. In this paper we give an implementation of the first phase of the force method which is numerically stable and preserves sparsity. A primary feature of our work is the development of an efficient algorithm for computing a banded basis for the null space by orthogonal decomposition. Numerical test comparisons for several practical structural analysis problems are provided.

Key Words: structural optimization, force method, orthogonal factorization, Givens rotations, turnback-QR method.

1. Introduction.

Given the external loads on a structure, the object of structural analysis is to determine the resulting internal forces, stresses, and displacements. The solution to this problem is provided by a variational principle (minimization of energy) subject to the linear elastic relationships among the nodes and elements of the finite element model of the structure. Either the forces or the displacements may be taken as the primary quantities to be computed, and the other can then be determined as a by-product. These two approaches give rise to the force (or flexibility) method and the displacement (or stiffness) method, respectively. In recent times the displacement method has predominated, largely because it is easier to implement on digital computers, especially for large sparse systems, and makes use of well established techniques of numerical linear algebra. The displacement method can be inefficient, however, for structural optimization problems in which a sequence of related structural analysis problems must be solved (e.g., problems having a fixed layout but differing material properties). The force method is then preferable because it utilizes a portion of earlier computations in order to solve such modified problems without starting the computations over from the beginning. Unfortunately, most

* Research sponsored by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy under contract W-7405-eng-26 with the Union Carbide Corporation, by the U.S.Army Research Office under contract no. DAAG29-81-k-0132, and by the Air Force under grant no. AFOSR-83-0255.

† Mathematics and Statistics Research Dept. Computer Sciences Div., Oak Ridge National Laboratory, Oak Ridge,Tennessee 37830.

‡ Depts.of Mathematics and Computer Science, North Carolina State Univ., Raleigh, N.C. 27650.

implementations of the force method have suffered from excessive fill or numerical difficulties, or both. In this paper we give an implementation of the first phase of the force method which is numerically stable and preserves sparsity for large-scale problems. The complete implementation will be published elsewhere [1].

Before stating our problem precisely, we need to develop some notation. The notational conventions used by structural engineers and by numerical analysts are quite different. As a compromise, we will use the same letters to denote various quantities as are commonly used by structural engineers (see, e.g., [3]), but we will retain the usual convention in numerical analysis that lower case letters represent vectors, while upper case letters represent matrices. Our notation is summarized in Table 1.

equilibrium matrix	:	E
element flexibility matrix	:	D
element stiffness matrix	:	D^{-1}
system stiffness matrix	:	$K=ED^{-1}E^{T}$
self-stress matrix	:	B
system flexibility matrix	:	$F=B^{T}DB$
nodal load vector	:	p
system force vector	:	f
nodal displacement vector	:	r
system displacement vector	:	v
redundant force vector	:	x

TABLE 1

The known, defining quantities for a structural analysis problem are the equilibrium matrix E, the element flexibility matrix D (or equivalently the element stiffness matrix D^{-1}), and the nodal load vector p. The unknowns to be determined are the system force vector f, the system displacement vector v, and the nodal displacement vector r. The remaining variables in Table 1, such as the redundant force vector x, are derived, intermediate quantities. We assume that E is an $m \times n$ matrix of rank m and that D is a symmetric positive definite matrix of order n. Here D is block diagonal with the diagonal blocks having orders ranging from one to six, depending on the finite element model of the structure. An example of a two-dimensional frame and its equilibrium matrix are shown in Figs. 1 and 2.

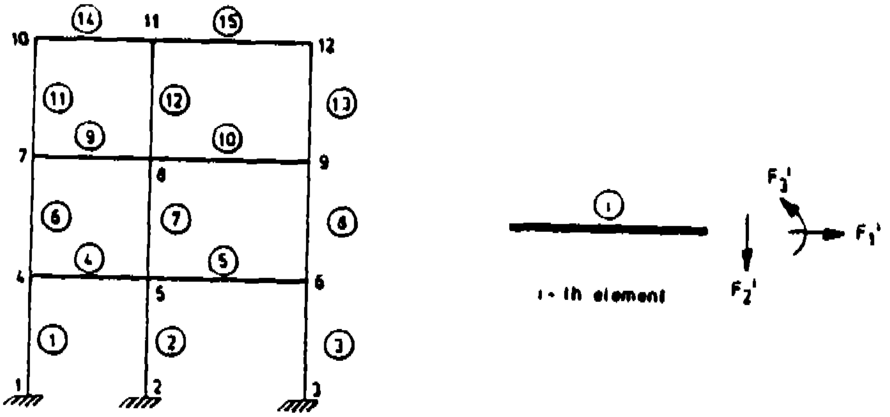


FIG. 1. *Two-dimensional frame with element and node numbering.*

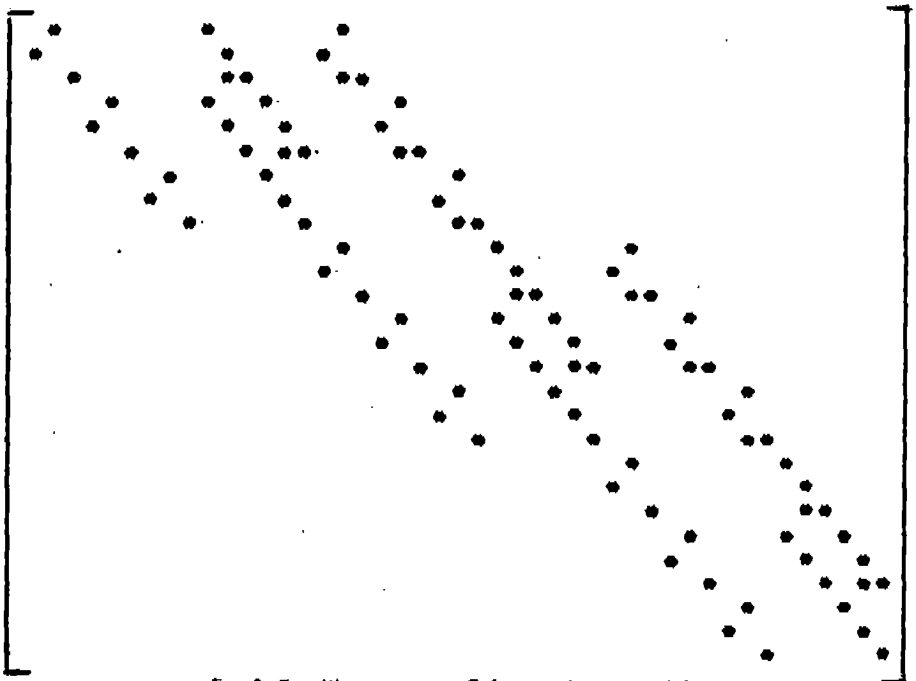


FIG. 2. *Equilibrium matrix E for two-dimensional frame.*

The linear elastic relationships for the system are

1. Equilibrium Equation:

$$Ef = p$$

2. Compatibility Equation:

$$E^T r = v$$

3. Material Equation:

$$Df = v.$$

The equilibrium and compatibility equations represent constraints on the forces and displacements, respectively, and the material equation represents the material characteristics.

The fundamental problem of structural analysis is that of solving the quadradic programming problem

$$\min_{f} \frac{1}{2} f^T D f \quad \text{subject to } Ef = p. \tag{1.1}$$

The force method is motivated by the observation that problem (1.1) is asking for a weighted, minimum-norm solution to the equilibrium equation. Given any particular solution s to such an underdetermined system, any other solution can be expressed as a sum of s and a solution to the corresponding homogeneous system Ef = 0. Thus we may write the system force vector f as

$$f = s + Bx, \tag{1.2}$$

where the n × (n - m) matrix B, called the self-stress matrix, is chosen so that its columns form a basis for the null space of E. Once s and B have been determined, we then need to determine the proper linear combination of the columns of B, represented by the vector x, so that f solves problem (1.1). The force methods is usually carried out in two phases:

Phase 1: Compute a particular solution s of the equilibrium equation, together with a self-stress matrix B such that EB = 0.

Phase 2: Compute the redundant force vector x and the system force vector f from (1.2).

Since Phase 1 depends only on E and p, it need be executed only once in order to solve a sequence of problems which differ only in D, with just Phase 2 being repreated for each new value of D. In order for the force method to be viable in practice, however, we need numerically stable algorithms for both phases which preserve sparsity. Phase 1 is the subject of this report. Phase 2 and its combination with Phase 1 will be considered elsewhere [1]. In the following section we describe a refined method for computing a banded basis matrix B for the null-space of E.

2. Band Schemes.

For a general sparse equilibrium matrix E it is difficult to say anything about the structure of the resulting self-stress matrix B other than that it might be quite full. If E has a banded structure, however, Topcu [4] and Kaneko, Lawo and Thierauf [2] have shown how by elimination methods to compute a B which also is banded. By modifying and extending their algorithm, we will show how by orthogonal methods to produce a banded self-stress matrix B. Since E and B are not square matrices, perhaps we should first point out that we do not mean banded with respect to the usual main diagonal, but with respect to a line from the upper left corner to the lower right corner of the matrix. Also, for our purposes the distinction between band and profile (also known as variable band, envelope, or skyline) implementations is unimportant, and so we will use the shorter term "band" to mean both possibilities.

We now give a refined turnback scheme, developed originally by Topcu [4] and by Kaneko, Lawo and Thierauf [2], for computing a banded basis matrix B for the null-space of a banded matrix A.

Given: An m × n banded matrix A, with full row rank m.

Purpose: To compute an $n \times \rho$, $\rho = n - m$, banded matrix B, with a small profile, whose columns form a basis of the null space of A; that is AB = 0, rank B = ρ .

Notation: $A^{(j)}$, $B^{(j)}$ denote the j^{th} columns of A and B respectively.

Step 1: Construct a sequence $b_1 < b_2 < \dots < b_\rho$, where b_j denotes the row index for the last nonzero entry in column $B^{(j)}$ of such a B, for $j = 1, \dots, \rho$.

Let A_m denote the first m columns of A. We determine the sequence of column indices $c_1 < c_2 < \dots < c_s$, if any, where column dependencies ("zero" pivots) occur when A_m is reduced to upper triangular form. Here either Gaussian elimination with row pivoting (or orthogonal reduction by Givens rotations) is applied to A_m. In particular, define, for $t = 1, \dots, s$,

$$c_t = \left\{\min\ k \;\middle|\; 1 \le k \le m \text{ and } \left\{A^{(1)}, \dots, A^{(k)}\right\} \setminus \left\{A^{(c_1)}, \dots, A^{(c_{t-1})}\right\} \text{ is a linearly dependent set.}\right\}$$

(Here $\left\{A^{(c_1)}, \dots, A^{(c_{t-1})}\right\}$ is missing if $t = 1$.)

Now we define the row indices b_j as follows. For $j = 1, \dots, \rho$,

$$b_j = \begin{cases} c_j \text{ if } 1 \le j \le s \\ m + j \text{ otherwise} \end{cases} .$$

Step 2: (Turnback step). For $j = 1, \dots, \rho$, consturct the column $B^{(j)}$ as follows:

(1) Set

$$t = \begin{cases} 1 \text{ if } b_j \le m \\ b_j - m \text{ if } b_j > m \end{cases}$$

and define a matrix E_j of columns from A as follows:

$$E_j = [E^{(b_j)} \mid E^{(b_j-1)} \mid, \dots, \mid E^{(t)}]$$

where the columns of E_j are columns of A, beginning with the b_j th column $E^{(b_j)} = A^{(b_j)}$, in decreasing column subscript order. More specifically the columns of E_j are the columr vectors

$$\left\{A^{(b_j)}, \ldots, A^{(t)}\right\} \Big\backslash \left\{A^{(t_1)}, \ldots, A^{(t_{j-1})}\right\}$$

in decreasing column subscript order, where $\left\{A^{(t_1)}, \ldots, A^{(t_{j-1})}\right\}$ is missing if j = 1, where the t's are given by (2.1) below.

(2) Apply either Gaussian elimination with row pivoting or orthogonal reduction by Givens rotations to E_j, producing an upper trapezoidal matrix

$$\hat{E}_j = [\hat{E}^{(b_j)} \quad \ldots \quad \hat{E}^{(t_j)}]$$

where

$$t_j = \max\left\{k \,\middle|\, k < b_j \text{ and } \left\{E^{(b_j)}, \ldots, E^{(t_j)}\right\} \text{ is a linearly dependent set.}\right\} \tag{2.1}$$

Then necessarily $\hat{E}_j$ is m × $(b_j - t_j)$ and has the form

$$\hat{E}_j = \begin{bmatrix} U & \hat{E}^{(t_j)} \\ 0 & 0 \end{bmatrix}$$

where U is square of order $b_j - t_j - 1$, upper triangular and nonsingular (the zero blocks may be missing).

(3) Solve the triangular system

$$Uy = -\hat{E}^{(t_j)}$$

by back substitution for $y = (y_s)$, $1 \leq s \leq b_j - t_j - 1$.

(4) Define the column $B^{(j)} = (B_i^{(j)})$ of B as follows

$$B_i^{(j)} = \begin{cases} 1 \text{ if } E^{(t_j)} \text{ is column } A^{(i)} \text{ of A} \\ y_s \text{ if } E^{(s)} \text{ is column } A^{(i)} \text{ of A, } 1 \leq s \leq b_j - t_j - 1 \\ 0 \text{ otherwise.} \end{cases}$$

□

3. Numerical Comparisons for Phase 1

3.1 Structural Test Data:

In addition to the two dimensional frame example described in section 1, five other structures provided a means for preliminary testing of the orthogonal turnback algorithm. Some preliminary comparisons were made between the orthogonal and the Gaussian elimination implementations of the algorithm. Each of the examples described below was kindly provided by M. Lawo of Essen, Germany.

Example 2:

Three Dimensional Frame.

The finite element model has 16 nodes and 24 elements and E is 72 × 144.

Example 3:

Plane Stress Problem (Slab).

The finite element model has 25 nodes and 16 elements and E is 40 × 80.

Example 4:

Plate Bending Problem.

The finite element model has 25 nodes and 16 elements and E is 59 × 144.

Example 5:

Plane Frame Problem (Wheel).

The finite element model has 32 nodes and 40 elements and E is 96 × 120.

Example 6:

Plane Stress Problem (Wrench).

The finite element model has 57 nodes and 48 elements and E is 112 × 216. A distinguishing feature here is a "hole" in the model. Also, three different node and element orderings were tested for this model: block angular, minimal bandwidth and minimal skyline.

3.2 Numerical Comparisons. Some preliminary comparisons were performed using each of the six structural analysis problems just described. In each example, the banded null-space basis matrix Z for the equilibrium matrix was computed by (1) the turnback algorithm implemented with Gaussian elimination and (2) the turnback algorithm implemented with orthogonal transformations by Givens rotations. The basic turnback algorithm in our refined form was given in Section 2.

All the software for this project was written at North Carolina State University in FORTRAN and executed on the large-scale IBM 3081 at the Triangle Universities Computing Center (TUCC). Each problem was tested in IBM double precision. IBM single precision was also used for certain of the problems in order to obtain a better basis for comparison. Various tolerances were tested for declaring a floating point zero pivot, i.e., a rank dependency among the columns of E. We settled on the tolerance 10^{-6} since the nonzero elements of E generally had magnitude approximately 1 and the data was given to six digits. This tolerance worked quite well in the computations.

We now give some observations on the performances of our software on these examples. In order to obtain some measure of the accuracy of a given method for computing a basis matrix B for the null-space of the equilibrium matrix E, the Frobenius norm, $||EB||_F$ was calculated. In what follows, B will denote a basis matrix computed by the back substitution method and Z will denote a basis matrix obtained by the refined turnback algorithm.

1. In each case $||EB||_F$ was smaller than $||EZ||_F$. In fact, the first norm was zero for the frame examples, Examples 1 and 2. Very little difference in accuracy was observed between the Gaussian elimination and orthogonal factorization implementations of the refined turnback method. Essentially,

$$||EZ||_F \doteq \begin{cases} 10^{-6} & \text{in single precision} \\ 10^{-15} & \text{in double precision} \end{cases}$$

2. Surprisingly, the orthogonal transformation implementation of the refined turnback algorithm executed faster than the Gaussian elimination implementation. We suspect that this situation is primarily due to the fact that row pivoting is used with Gaussian elimination while no pivoting is necessary for stability with Givens rotations. Such pivoting can consume an inordinate amount of time in the lengthy turnback phase of the algorithm.

The computations we have performed thus far are preliminary in nature, but on the basis of the results discussed above, the orthogonal factorization implementation of the refined turnback algorithm appears to be an attractive method for computing a banded basis for the null-space.

Further refinements of these methods followed by numerical tests will be made in the near future under grants from other funding agencies.

REFERENCES

[1] M. T. Heath, R. J. Plemmons, and R. C. Ward, Sparse orthogonal schemes for structural optimization using the force method, ORNL Tech. Report. no. CSD-88, Oak Ridge, TN, May, 1983.

[2] I. Kaneko., M. Lawo and G. Thierauf, On computational procedures for the force method, Int. J. Numer. Meth. Engrg., 18(1982), pp. 1469-1495.

[3] J. Robinson, Integrated Theory of Finite Element Methods, John Wiley and Sons, New York, 1973.

[4] A. Topcu, A contribution to the symmetric analysis of finite element structures using the force method (in German), Doctoral Thesis, University of Essen, Germany, 1979.

THREE DIMENSIONAL STRESS ANALYSIS OF AN ARTILLERY PROJECTILE JOINT

Tien-Yu Tsui
Army Materials and Mechanics Research Center
Watertown, Ma. 02172

M.L. Chiesa and M.L. Callabresi
Sandia National Laboratories
Livermore, CA. 94550

ABSTRACT

This study presents the results of a three dimensional elastic-plastic dynamic stress analysis of one of the structural joints encountered in artillery projectiles. The particular spline joint analyzed has equally spaced set screws around the surface of the projectile. In general, these types of structural joint problems are variable contact problems in that the interaction between the set screws and their bearing surface along with the interaction between the interfaces of the joint are nonlinear in nature. Due to the complexity of the structural configuration and loadings of the joint, the finite element method has been used to solve the problem. The numerical analysis covers the time from initial launching to barrel exit. Stresses and deformations in the joint are determined at various stages of loading. The effect of the set screws and set screw holes on the stress distributions in the joint is examined in detail.

INTRODUCTION:

Artillery projectiles are subjected to extremely high loads during firing. At present, the design of artillery projectiles is greatly facilitated by the exploitation of the finite element method. But the application of the method has been limited to the simplified two dimensional or axisymmetric analysis. This is due to the complexity of the geometries of the projectiles thus requiring very long computer analysis. Although successes were achieved in the previous designs, there have always been concerns of the structural integrity of the projectiles with the presence of high local stresses. In the case of XM753 projectiles, high local stresses appear in the region of the pinned joint. In the case of XM785 projectiles the use of spline joint with set screws in the design also leads to high local stresses in the joint area.

After firing, an artillery projectile is subjected to various loads continuously changing with time. It experiences first a very high compressive load in the axial direction during in bore flight and then a high tensile load at barrel exit. Such a loading history causes the projectile to undergo a stress reversal i.e. from a stress state of compression to that of tension. However it has been a common practice in the design of artillery projectiles to perform either one or two simple independent stress analysis. The one which is generally carried out is a two dimensional quasi-static analysis in which the loads at the peak linear acceleration during in bore flight are used to compute the stresses and deformations in the projectile.

Sometimes a two dimensional dynamic analysis is also conducted to determine the stresses and deformations in the projectile caused by the sudden drop off of pressure of the propellant gas at barrel exit.

In the present investigation a detailed three dimensional elastic-plastic dynamic stress analysis of a spline joint of XM785 projectile is performed. The purposes of the study are:

(1) to verify the structural adequacy of a proposed design.

(2) to asses the effect of set screws and set screw holes on the stress distributions in the joint region of the projectile.

In addition the present work also serves as an initial effort to determine the extent of the influence of the stress reversal or Bauschinger effect on the structural integrity of the projectile.

DESCRIPTION OF LOADING CONDITIONS

After a projectile is fired and before it departs the gun barrel it is subjected to a combination of the following loads:

(1) axial compressive load due to linear acceleration of the projectile

(2) centrifugal load due to angular rotation of the projectile.

(3) torsional load due to angular acceleration of the projectile.

(4) internal load due to interaction of interior components and projectile.

(5) external load due to gun tube constraint, rotating band pressure and balloting.

As the projectile departs the gun barrel, it experiences a tensile load or a negative set-back load (elastic release) in the axial direction resulting from the sudden drop off of propellant pressure at the barrel exit. Among the loads the axial load is the dominating one.

It has been a general practice to omit the effect of the torsional load induced by the angular acceleration in the design of the projectiles. Analysis has shown that as a result of such an omission, the magnitude of the effective stresses in the projectiles are about 2% - 4% lower (1). Since the present investigation concerns the determination of the stress distributions in the region of the joint, only the portion of the projectile in the neighborhood of the joint is considered. The area where the rotating band is located is excluded. Thus the load due to rotating band pressure is not included in the analysis. Only the axial, centrifugal and internal loads during in bore flight and the negative set back load at barrel exit are considered in the three dimensional dynamic stress analysis of the joint. Fig. 1 shows the linear acceleration of the projectile used in this analysis for the calculation of the axial load. The projectile reaches an acceleration of 17,000g in about 6 milliseconds, zero acceleration at barrel exit and immediately is subjected to a negative acceleration of 2000g.

METHOD OF STRESS ANALYSIS

Figs. 2 and 3 shows respectively the geometric configuration of an artillery projectile and a typical section of its joint. The latter is formed by passing a plane at the midpoint of two neighboring set screws and another plane through the center of one of the set screws. The two planes are parallel to the axis of the projectile.

In the initial phase of the analysis, the finite element model of the joint (see Fig. 4) was created by using the computer program PATRAN-G (2). It employs sophisticated interactive color graphics and a powerful geometry-based language for geometry construction and finite element modeling during the first phase of the program. Then nodal point and finite element generation, assignment of physical properties, application of external loads and definition of sliding interfaces are accomplished in the second phase of the program.

The explicit three dimensional finite element code DYNA3D (3) was used to compute the stresses and deformations in the finite element model of the joint. DYNA3D is designed to analyze the large deformation dynamic response of inelastic solids. It has a contact algorithm that can model gaps and sliding materials interfaces. It uses a 8-node constant stress solid element and one point integration in element stiffness calculations. It is programmed to take full advantage of vector optimization on the CRAY-1 (a class VI machine) and can execute at less than 0.67CPU (central processor Units) minutes per million mesh cycles. A symmetric, penalty based, contact-impact algorithm was implemented that not only reduced hourglassing problems, but also was considerably faster in execution speed and exceedingly reliable.

There were 1773 eight node brick elements and 2611 nodes in the finite element model of the joint. Four sets of gap or sliding interfaces were required in the model. Three sets of sliding interfaces were used to model the contact between the case structure and the wedge. A sliding interface with small initial gap was used to model the interaction between the set screw and the hole.

The computations were performed on a CRAY-1 computer at Sandia National Laboratories. The computer time required to complete the dynamic analysis was about 2.8 CPU hours.

DISCUSSION OF RESULTS

The post processor program GRAPE (4) was used to obtain plots of stress contours on the surfaces of the joint and the deformed shapes of the joint at the time when the maximum linear acceleration is reached and also at barrel exit. These are shown in Figs. 5 to 11. Examination of the results of the analysis leads to the following findings:

(1) Due to the presence of set screws, a larger portion of the axial compressive load is transmitted, during in-bore flight, through the center sections of the regions bounded by each pair of set screws. Consequently, the center

sections experience higher stresses. This is different from the uniform stress distribution in the circumferential direction found in the two dimensional stress analysis.

2. Higher stresses occur on the outer surfaces of the case structure and the wedge at both peak axial compression and tension at barrel exit due to bending effect.

3. Local yielding occurs in the set screws, set screw holes and other areas in the case structure and the wedge at time of peak axial compression and that of tension at barrel exit. Initial examination of the plastic strains (or Bauschinger effect) at peak axial compression leads one to believe that they have a negligible effect on the level of stresses in the joint at barrel exit. This is because their magnitudes are small (.001) and their locations are also different from the plastic strains induced by the tensile load or negative set-back at barrel exit.

4. Similar stress distributions were obtained when a wedge made of either steel or titanium was used with a titanium case structure.

In summary, a detailed three dimensional elastic-plastic dynamic stress analysis of an artillery projectile was obtained to determine the adequacy of a proposed design. The analysis was performed by using the DYNA3D finite element program. The results of the present analysis will be verified by laboratory test currently under preparation at AMMRC. Future analytical work will include application of fracture mechanics to the problem and accurate determination of the Bauschinger effect by using actual stress strain curves of the materials. The present analysis employed isotropic strain hardening which does not accurately represent the material behavior in a loading condition where there is a reversal of applied loads.

ACKNOWLEDGEMENT

The support and encouragement of Mr. C. Spinelli, Development and Management Officer for XM785 projectile, Nuclear Projectile Branch at ARRADCOM, Dover, N.J. and Dr. G.A. Benedetti, Chief, Applied Mechanics Division, Sandia National Laboratories, are gratefully acknowledged.

REFERENCES

1. T. Tsui, "On The Magnitude Of Stresses Due To Angular Acceleration Of Projectile", unpublished notes.

2. PATRAN-G, PDA Engineering, 1560 Brookhollow Drive, Santa Ana, California 92705

3. J.O. Hallquist, "DYNA3D-Nonlinear Dynamic Analysis Of Solids In Three Dimensions", Lawrence Livermore, Rept. UCID-19592, Nov. 1982.

4. B.E. Brown, "Displacing The Results Of Three Dimensional Analysis Using GRAPE", Lawrence Livermore National Laboratory, Rept. UCID-18507, Oct. 1979.

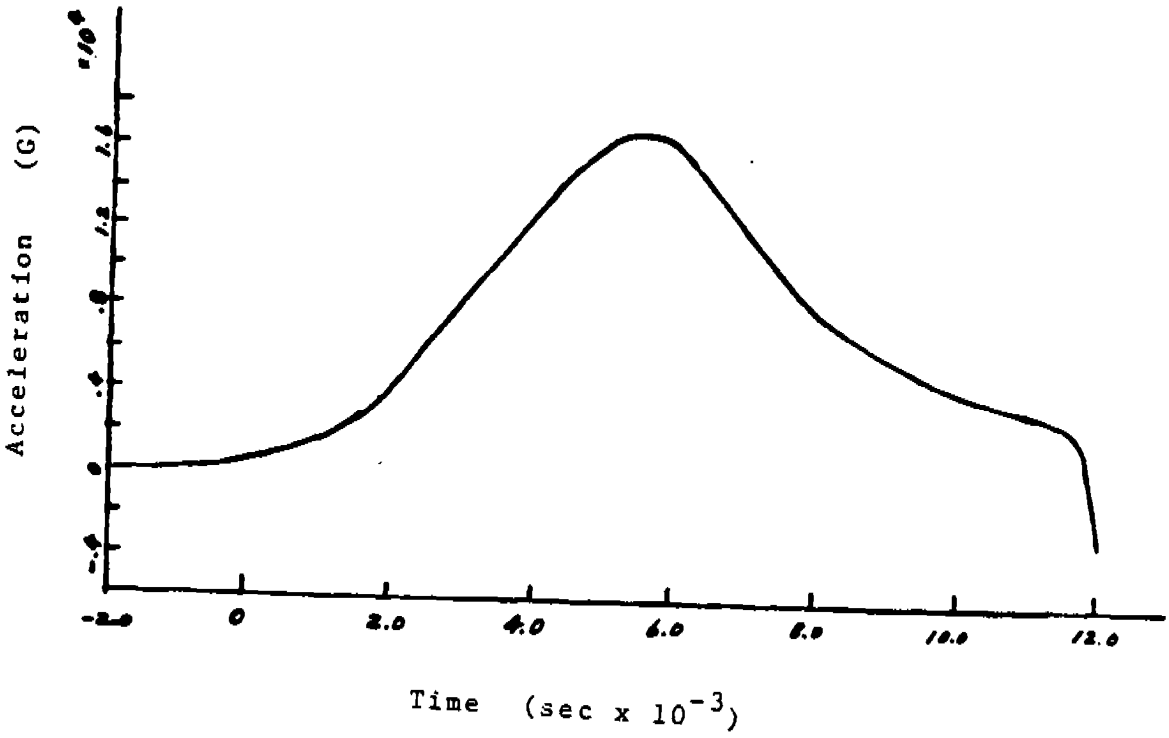


Fig. 1 Linear Acceleration of Projectile

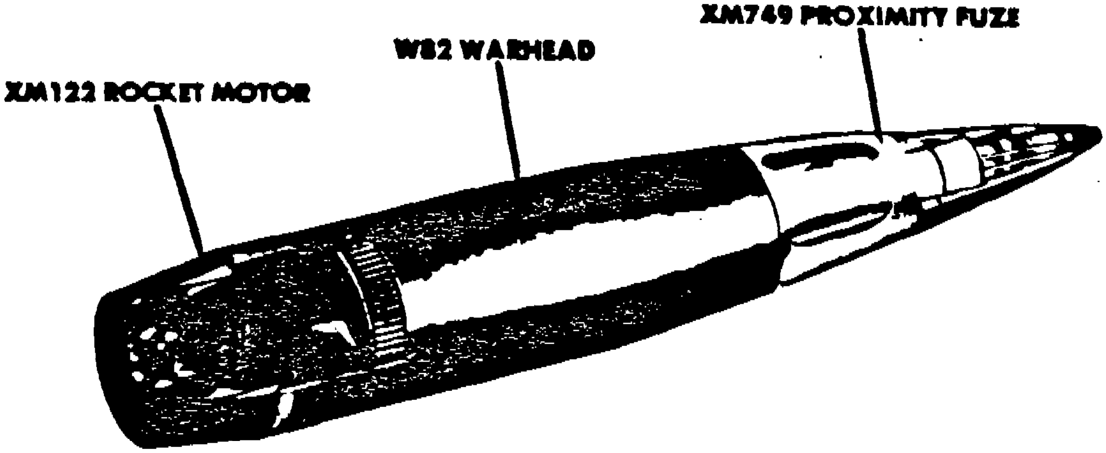


Fig. 2 Projectile, Atomic, 155MM: XM785

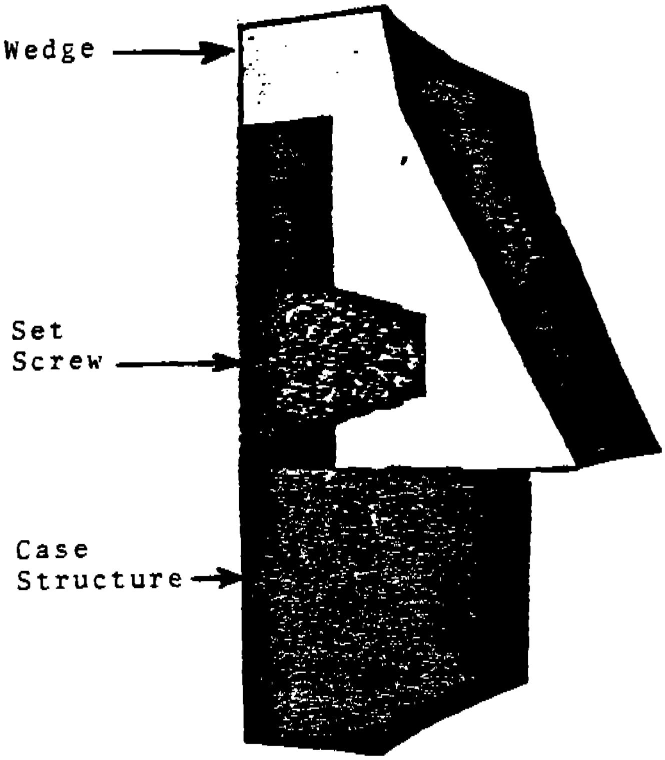


Fig. 3 Typical Joint Section

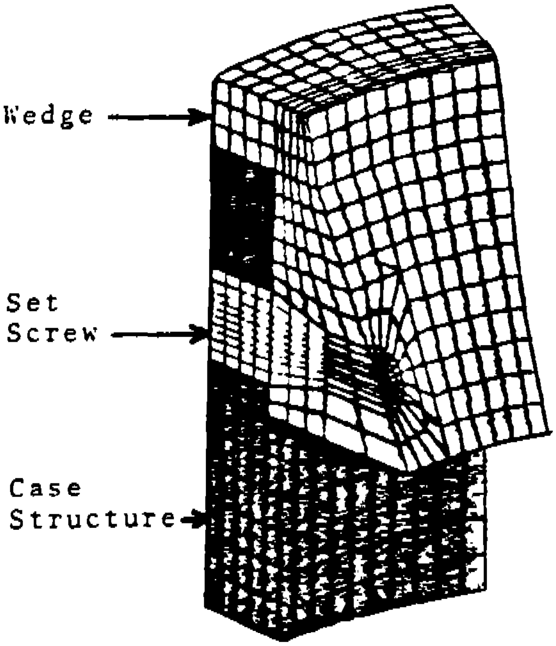
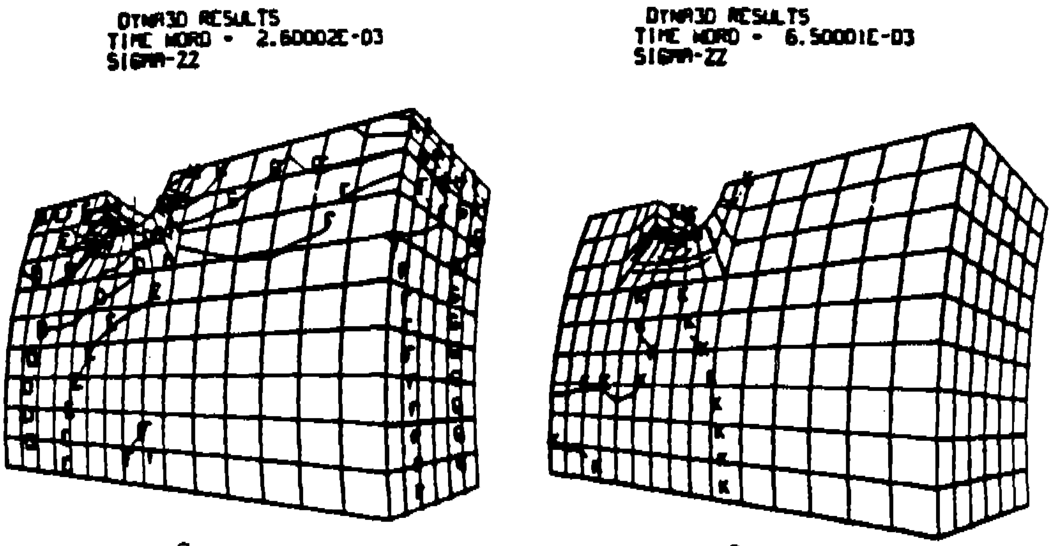


Fig. 4 Finite Element Model

See Table 1 for Values of Symbols



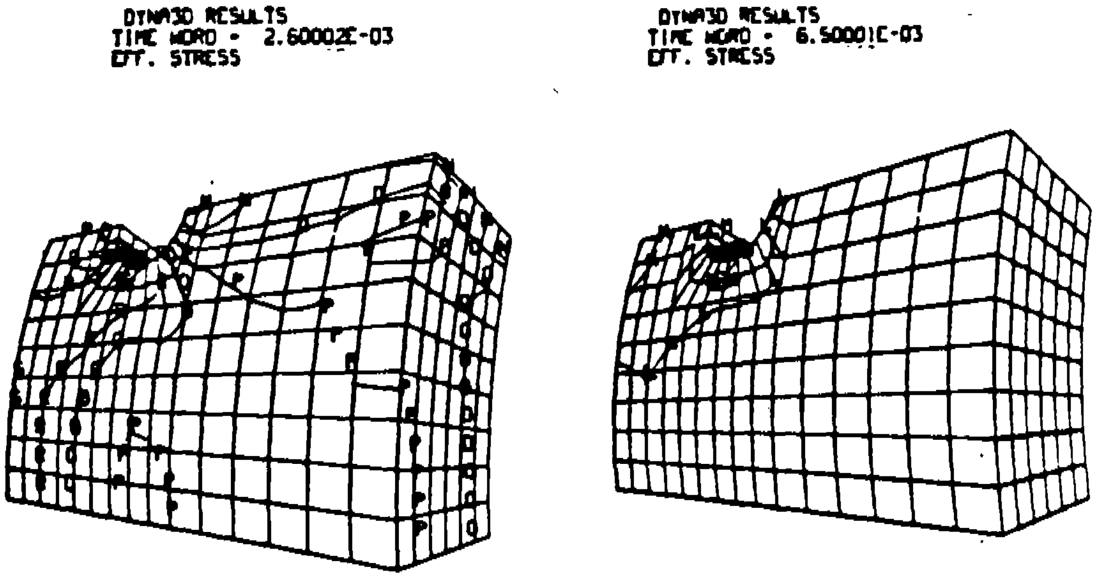
Case 1

Case 2

Fig. 5 Axial Stresses in the Case Structure

Case 1: At time of maximum compression

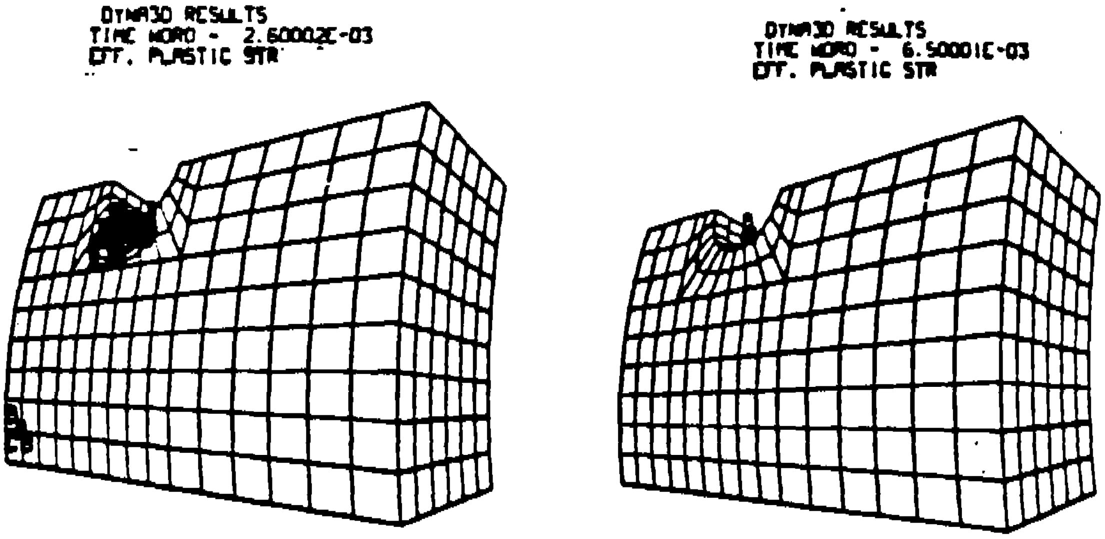
Case 2: At time of barrel exit



Case 1

Case 2

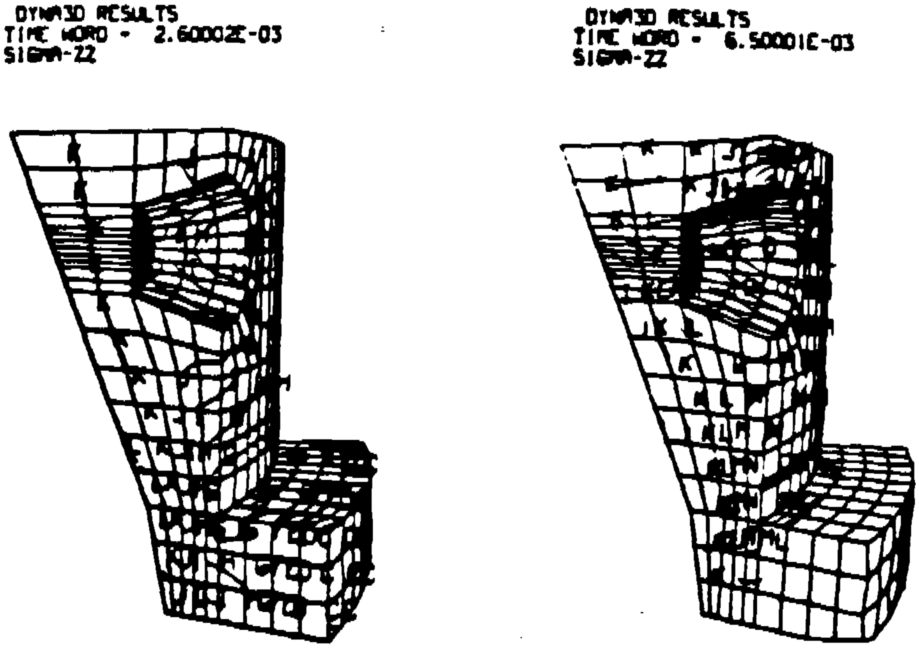
Fig. 6 Effective Stresses in the Case Structure



Case 1

Case 2

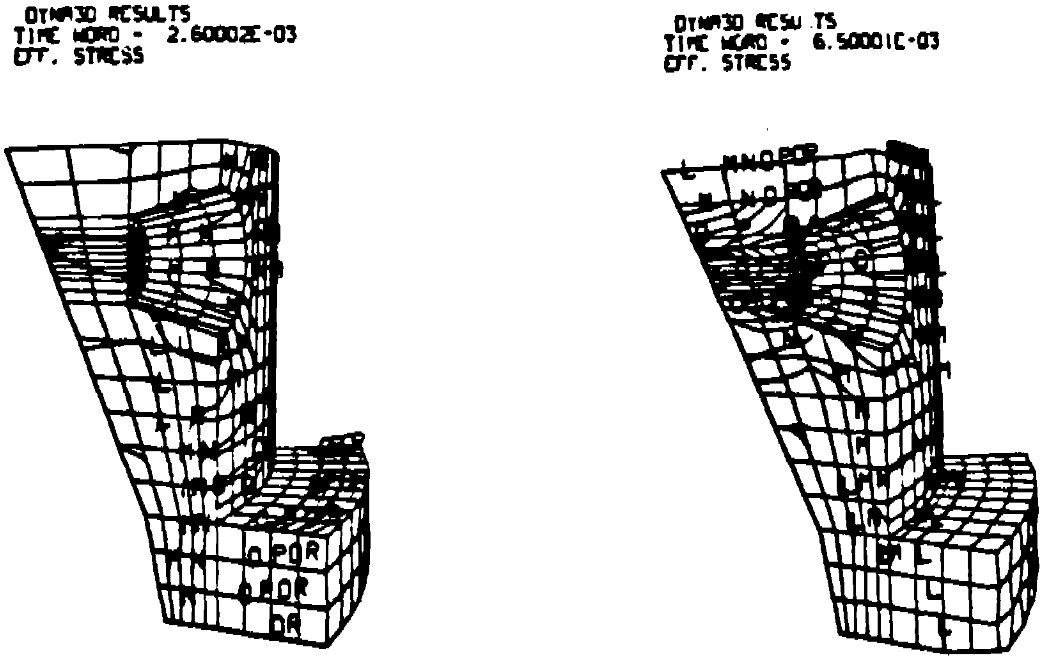
Fig. 7 Effective Plastic Strains in the Case Structure



Case 1

Case 2

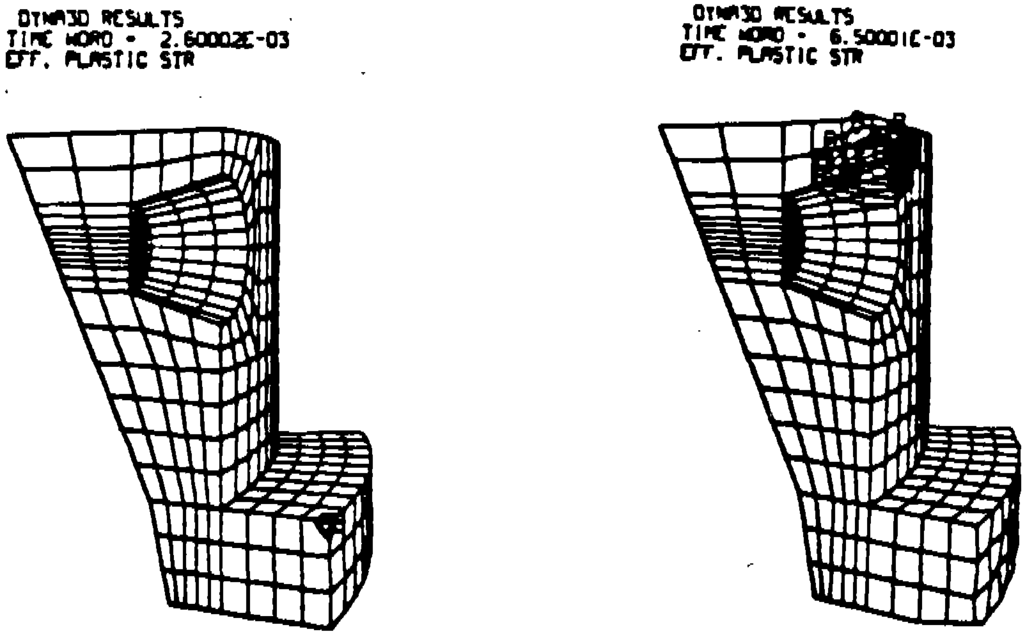
Fig. 8 Axial Stresses in the Wedge



Case 1

Case 2

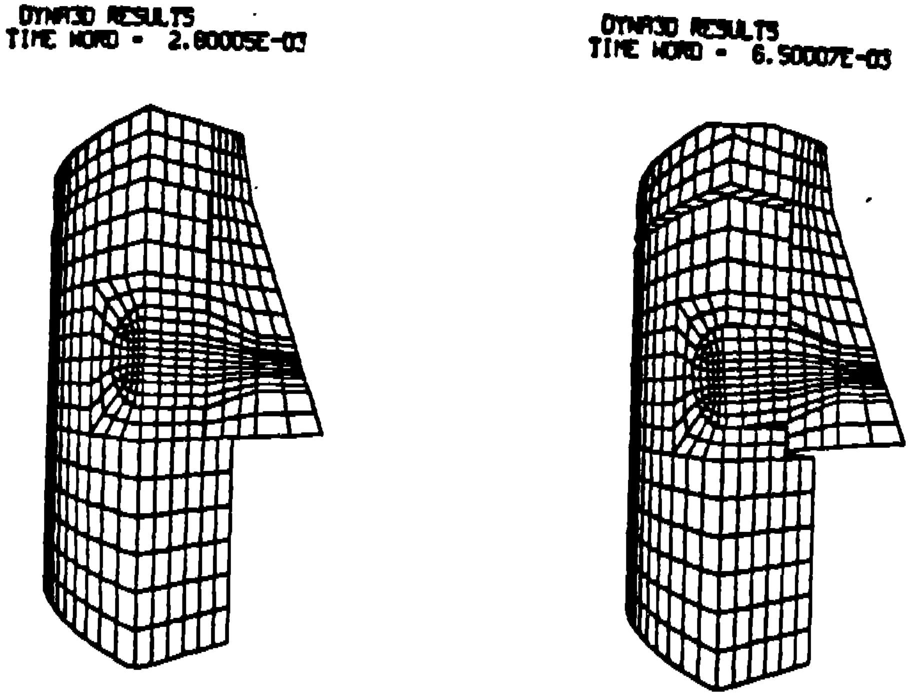
Fig. 9 Effective Stresses in the Wedge



Case 1

Case 2

Fig. 10 Effective Plastic Strains in the Wedge



Case 1

Case 2

Fig. 11 Deformed Shape of Finite Element Model

Table 1 - Levels Of Stresses and Strains

	Stress (ksi)	Strain (case 1)	Strain (case 2)
A	-200	0	0
B	-180	.001	.01
C	-160	.002	.02
D	-140	.003	.03
E	-120		
F	-100		
G	-80		
H	-60		
I	-40		
J	-20		
K	0		
L	20		
M	40		
N	60		
O	80		
P	100		
Q	120		
R	140		
S	160		
T	180		
U	200		

A SIMPLE APPROACH FOR DETERMINATION OF BURSTING PRESSURE OF A THICK-WALLED CYLINDER

SHIH C. CHU

TECHNOLOGY BRANCH, ARMAMENT DIVISION
FIRE CONTROL & SMALL CALIBER WEAPON SYSTEMS LABORATORY
U.S. ARMY ARMAMENT RESEARCH AND DEVELOPMENT COMMAND, DOVER, NJ 07801

ABSTRACT

By using the true-stress true-strain relation, a simple finite-strain incompressible analytical solution technique is developed to predict pressure-deformation relations up to the failure of thick-walled cylinders subjected to internal pressure. The material is assumed to be an isotropic hardening material that obeys the von Mises yield condition. The flow law for the material incorporates the Prandtl-Reuss stress-strain relations and a loading function represented by the true-stress versus true-strain diagram. Poisson's ratio is assumed to be equal to one-half for both elastic and plastic strains. The bursting pressure for thick-walled cylinders made of either SAE 1045 steel or copper is presented graphically for various ratios of outer and inner radii of undeformed cylinders. The true location of inner surface of cylinders made of both metals at various pressure levels was determined and presented graphically. Pressure-expansion curves for cylinders made of SAE 1045 steel and OFHC copper are presented.

NOMENCLATURE

σ_r, σ_θ, σ_z	true-stress components
ε_r, ε_θ, ε_z	true-strain components
σ_e, ε_e	effecive true stress and effective true strain
r_{01}, r_{02}	inner and outer radii of the undeformed thick-walled cylinder
r_1, r_2	inner and outer radii of the deformed thick-walled cylinder
r	variable radius of deformed thick-walled cylinder
r_o	variable radius of undeformed thick-walled cylinder
u	$r-r_o$ is the radial displacement
P_1, P_2	internal and external pressures
E	Young's modulus
ν	Poisson's ratio
σ_o	σ_{o1} is yield stress
σ_{oi}	stress at intersection of equations 3 and 4
ε_o	σ_o/E is yield strain
K	ratio of ε_e to ε_o at $r=r_1$
β	ratio of ε_z to $\varepsilon_{e(r=r_1)}$
ρ	r_{02}/r_{01} radius ratio
f_u	ultimate tensile stress
P_u	bursting pressure

INTRODUCTION

Thick-walled cylinders, such as gun tubes and pipes to hydraulic presses, often are required to resist internal pressures that will cause inelastic strains. It is important for the engineer to have a relatively reliable and simple solution technique that can be used to predict the load-deformation relations for thick-walled cylinders subjected to internal pressure up to fracture load encountered. This information would enable the engineer to determine the factor of safety of his design against failure by fracture. In the case of autofrettage, it would enable the engineer to determine safe pressures to be used or to determine safe limits for the deformations. Partially plastic, thick-walled cylinders have been investigated by many investigators [1-6]. In all the solutions [1-6], it has been assumed that subsequent to the initial yield the elastic plastic interface is cylindrical and concentric with the bore. This assumption is correct in many cases but not correct for a material with a large drop of stress at yield. In addition most of the solutions are based on the total strain theory, though Taylor has pointed out that the incremental theory should be used. However, the total strain theory gives a solution which is only slightly in error in this particular problem. The governing equations have been derived for partial yield thick-walled cylinders, assuming the dimensional changes are negligible, but in the region of the bursting pressure the strains are considerably large and this assumption is no longer valid.

Based on the closed-ended cylinder condition, bursting pressure was investigated by Manning and Chem [7] and Crossland and Bones [8]. In their approach the plastic strain in the axial direction is assumed to be zero. Experimental data for closed-ended cylinders [9] indicates that the assumption is reasonable.

Some empirical equations for bursting pressure of thick-walled cylinders have been proposed by Faupel and Furbeck [10] and Iterson [11]. Those empirical equations are derived either based on experimental data or on intuition. The well-known mean-diameter formula is

$$P_u = 2f_u \frac{(\rho-1)}{(\rho+1)} \qquad (1)$$

A finite total-strain incompressible analytical solution is presented for thick-walled cylinders subjected to internal pressure. Using the strain calculations, the geometry of the deformed cylinder is calculated at the end of each load increment. Since the deformed geometry of the cylinder is used in making the calculations of the strain increments for each load increment, the computed strains are true strains and the solution is valid for finite strain. One might question the advisability of considering a solution not readily acceptable to experts in the field of plasticity when an incremental compressible solution is available. The total-strain solution was used because it was inexpensive to use and the computer program did not have convergence problems. On the other hand, the computer program for the incremental solution would not converge unless the plastic-strain increment for each increment of load was small. In order to obtain a compressible incremental solution it was necessary to make the load increments smaller and smaller as the plastic strains increased in magnitude. Because of the convergence problems and the excessive computer cost for implementing the compressible incremental theory, the present solution technique is presented.

GOVERNING EQUATIONS

In this investigation, the material is assumed to be an isotropic hardening material that obeys the von Mises yield condition. Poisson's ratio is assumed to be equal to one-half for both elastic and plastic strains.

Cylindrical coordinates are used with z being the axial direction and r and θ being the radial and circumferential directions, respectively. The nonzero stress and strain components are σ_r, σ_θ, σ_z, $\tau_{\theta z}$, ε_r, ε_θ, ε_z and, $\gamma_{\theta z}$. For the thick-walled cylinder subjected to pressure, the effective stress σ_e and effective strain ε_e are defined by the relations

$$\sigma_e = \frac{1}{\sqrt{2}}\sqrt{(\sigma_r-\sigma_\theta)^2 + (\sigma_\theta-\sigma_z)^2 + (\sigma_z-\sigma_r)^2} \tag{2}$$

$$\varepsilon_e = \frac{\sqrt{2}}{3}\sqrt{(\varepsilon_r-\varepsilon_\theta)^2 + (\varepsilon_\theta-\varepsilon_z)^2 + (\varepsilon_z-\varepsilon_r)^2} \tag{3}$$

The flow law for the material incorporates the Prandtl-Reuss stress-strain relations and a loading function represented by a tension true-stress versus true-strain diagram. The loading function for the analytical solutions is obtained from a tension specimen whose effective true-stress (σ_e) versus effective true-strain (ε_e) diagram is approximated by a finite number of straight lines. The straight line through the origin is given by the relation

$$\sigma_e = E\varepsilon_e \tag{4}$$

where E is Young's modulus. All of the straight lines are given by the relation

$$\sigma_e = (1-m_i)\sigma_{oi} + m_i E \varepsilon_e \tag{5}$$

where σ_{oi} is the stress at the intersection of the two straight lines given by eqs. (4) and (5) and $m_i E$ is the slope of the straight line given by eq. (5).

The condition of incompressibility for the cylinder is given by

$$\varepsilon_r + \varepsilon_\theta + \varepsilon_z = 0 \tag{6}$$

The equation of equilibrium for the thick-walled cylinder is given by the relation

$$r \frac{d\sigma_r}{dr} = \sigma_\theta - \sigma_r \tag{7}$$

where r is the variable radius for the deformed cylinder. A point at radius r_o in the undeformed cylinder has a radial displacement u during loading of the cylinder so that $r = r_o + u$. The true magnitude of the radial and circumferential strains for incompressible conditions are given by the relations

$$\varepsilon_r = \ell n\left(1 + \frac{du}{dr_o}\right) \tag{8}$$

and

$$\varepsilon_\theta = \ell n\left(1 + \frac{u}{r_o}\right) \tag{9}$$

The compatibility condition for the thick-walled cylinder can be obtained by taking the derivative of eq. (9) with respect to r_o

$$r_o e^{\varepsilon_\theta} \frac{d\varepsilon_\theta}{dr_o} = \frac{du}{dr_o} - \frac{u}{r_o} \tag{10}$$

Equation (10) can be simplified by using eqs. (6), (8), and (9) to give

$$r_o \frac{d\varepsilon_\theta}{dr_o} = \frac{e^{-2\varepsilon_\theta}}{e^{\varepsilon_z}} -1 \tag{11}$$

where e^{ε_z} is independent of r_o and is assumed to be known.

Based on Hencky's stress-strain relations:

$$\frac{\varepsilon_r-\varepsilon_\theta}{\sigma_r-\sigma_\theta} = \frac{\varepsilon_\theta-\varepsilon_z}{\sigma_\theta-\sigma_z} = \frac{\varepsilon_z-\varepsilon_r}{\sigma_z-\sigma_r} = \frac{3\varepsilon_e}{2\sigma_e} \tag{12}$$

From eq. (12) one obtains

$$\varepsilon_z = \frac{3\varepsilon_e}{2\sigma_e} (\sigma_z-\sigma_r) + \varepsilon_r \tag{13}$$

By using the condition of incompressibility, eqs. (6) and (13), one can find

$$\varepsilon_r = \frac{\varepsilon_e}{\sigma_e} [\sigma_r - \frac{1}{2} (\sigma_\theta+\sigma_z)] \tag{14}$$

Similarly,

$$\varepsilon_\theta = \frac{\varepsilon_e}{\sigma_e} [\sigma_\theta - \frac{1}{2} (\sigma_z+\sigma_r)] \tag{15}$$

$$\varepsilon_z = \frac{\varepsilon_e}{\sigma_e} [\sigma_z - \frac{1}{2} (\sigma_r+\sigma_\theta)] \tag{16}$$

The problem considered here is a thick-walled cylinder subjected to internal pressure only. The computer solution for the total-strain, incompressible analytical solution is obtained by specifying the deformations for the thick-walled cylinder and calculating the pressure. This is done by specifying values for the true strains ε_θ and ε_z at the inner radius of the deformed cylinder. The effective true strain is specified as

$$\varepsilon_{e(r=r_1)} = K\varepsilon_o \qquad (17)$$

where $\varepsilon_o = \sigma_{01}/E$ and σ_{01} is the value of σ_{oi} in eq. (5) when $i = 1$, σ_{01} is assumed to be the elastic limit of the material. The axial strain ε_z is given by specifying values for β in the relation

$$\varepsilon_z = \beta K\varepsilon_o \qquad (18)$$

The solution of the differential equation, eq. (11), can be found in the form

$$\varepsilon_\theta = \frac{1}{2} \ln \left(\frac{1}{e^{\varepsilon_z}} + \frac{C}{r_o^2}\right) \qquad (19)$$

where C is the constant of integration which is obtained from the fact that $\varepsilon_\theta = \varepsilon_{\theta_1}$ is known when $r = r_1$ $(r_o = r_{01})$, thus

$$e^{2\varepsilon_{\theta_1}} = \frac{r_1^2}{r_{01}^2} = \frac{1}{e^{\varepsilon_z}} + \frac{C}{r_{01}^2} \qquad (20)$$

solving for C

$$C = \frac{r_1^2 e^{\varepsilon_z} - r_{01}^2}{e^{\varepsilon_z}} \qquad (21)$$

Substituting C into eq. (19) gives

$$\varepsilon_\theta = \frac{1}{2} \ln \frac{r_o^2 - r_{01}^2 + r_1^2 e^{\varepsilon_z}}{r_o^2 e^{\varepsilon_z}} \tag{22}$$

If ε_θ is eliminated by using eq. (6)

$$\varepsilon_r = - \frac{1}{2} \ln \left(\frac{1}{e^{\beta K \varepsilon_o}} + \frac{C}{r_o^2}\right) - \beta K \varepsilon_o \tag{23}$$

The effective strain may be obtained from eq. (3). By eliminating ε_r, one has

$$\varepsilon_e = \frac{2}{\sqrt{3}} \sqrt{\varepsilon_\theta{}^2 + \varepsilon_\theta \varepsilon_z + \varepsilon_z^2} \tag{24}$$

or

$$\varepsilon_\theta{}^2 + \varepsilon_\theta \varepsilon_z + \varepsilon_z{}^2 = \frac{3}{4} \varepsilon_e{}^2 \tag{25}$$

At the inner surface, ε_z and ε_e are specified; hence, one can solve for $\varepsilon_\theta = \varepsilon_{\theta 1}$. By using eq. (20), r_1 can be determined.

The stress distributions in the deformed thick-walled cylinder can be determined as follows. Substituting eq. (18) into eq. (16) one obtains

$$\sigma_z = \frac{\sigma_e}{\varepsilon_e} \beta K \varepsilon_o + \frac{1}{2} (\sigma_\theta + \sigma_r) \tag{26}$$

Equation (2) and eq. (26) reduces to

$$\sigma_\theta - \sigma_r = \frac{2}{\sqrt{3}} \sigma_e \sqrt{1 - \left(\frac{\beta K \varepsilon_o}{\varepsilon_e}\right)^2} \tag{27}$$

which can be substituted into the equation of equilibrium, eq. (7), to give

$$r \frac{d\sigma_r}{dr} = \frac{2}{\sqrt{3}} \sigma_e \sqrt{1-(\frac{\beta K \varepsilon_o}{\varepsilon_e})^2} \qquad (28)$$

Before eq. (28) can be integrated, it is necessary that σ_e and ε_e be expressed as functions of r. This can be carried out by the following computational procedure.

COMPUTATIONAL PROCEDURE

The undeformed thick-walled cylinder is divided into N rings of equal thickness. For each specified value (monotone increase) of β and K, the following computation will be performed:

Step 1. Calculate $\varepsilon_e(r=r_1)$ and ε_z, by using eqs. (17) and (18).

Step 2. Solve $\varepsilon_{\theta 1}$, by substituting ε_e and ε_z into eq. (25).

Step 3. Calculate r_1 by using eq. (20).

Step 4. Calculate C, by using eq. (21).

Step 5. Calculate ε_θ, ε_z, and ε_r at each ring stations, by using eqs. (22), (18), and (23), respectively.

Step 6. The deformed position of each ring can readily be determined by using the following relation

$$r = r_o e^{\varepsilon_\theta} \qquad (29)$$

Step 7. The effective strain ε_e at each ring station in the deformed cylinder can be determined by using eq. (3).

Step 8. The loading function represented by eqs. (4) and (5) is used to calculate the effective stress σ_e at each ring station.

Step 9. Starting with the known radial stress $\sigma_{r_2} = -P_2$ at the outer radius, the radial stress σ_r is obtained by numerical integration of eq. (28).

Step 10. Substituting σ_r into eq. (27) σ_θ can be calculated.

Step 11. Calculate σ_z by substituting σ_θ and σ_r into eq. (26).

Step 12. For each specified value of β, the loads on the thick-walled cylinder are calculated for increasing values of K. Calculations are stopped when the current value of internal pressure is smaller than the value of internal pressure for the previous specified value of K. The bursting pressure is defined as the internal pressure P_1.

COMPUTATIONAL RESULTS

Based on experimental testing data, true-stress versus true-strain relations for the two metals, steel and copper, are shown in Fig. 1. The plastic portion of the true-stress versus true-strian diagram is approximated by five straight lines for the SAE 1045 steel and by seven straight lines for OFHC copper. Each straight line is represented by eq.(5); values of σ_0 and m for each straight line are listed in Table 1. Each true-stress versus true-strain diagram can be approximated by a finite number of straight lines with extreme accuracy. If the straight lines are chosen so that the actual diagram lies above and below the staight-line segment approximation with the difference no more than one percent (1%), the error introduced by the apprxomation will be a fraction of 1%.

Table 1. Values of m and σ_0 for Straight Lines Approximating Plastic Portion of Stress-Strain Diagram

Straight Line	σ_0 ksi	σ_0 MPa	m
	SAE 1045 Steel		
1	43.4	299	0.05083
2	54.0	372	0.02858
3	80.0	552	0.00847
4	95.0	655	0.00309
5	111.0	765	0.00128
	OFHC Copper		
1	2.50	17.2	0.17125
2	3.25	22.4	0.07063
3	4.00	27.6	0.03125
4	5.37	37.0	0.01991
5	8.40	57.9	0.01313
6	21.0	144.8	0.00450
7	39.0	269.9	0.00078

The bursting pressures for thick-walled cylinders made of SAE 1045 steel and OFHC copper for various radius ratios (ρ = 1.25, 1.5, . . ., 3.75 and 4.00) were calculated based on the computational procedure proposed in the previous section. The resulting bursting pressure is plotted against the radius ratio as shown in Figs. 2 and 3.

The pressure-expansion curves for cylinders made of SAE 1045 steel and OFHC copper are shown in Figs. 4 and 5, respectively. Each curve is stopped when the bursting pressure is encountered.

The displacements of the inner surface for thick-walled cylinders made of SAE 1045 steel and OFHC copper were calculated by using eq. (20) and are shown in Figs. 6 and 7, respectively.

CONCLUSIONS

By using the true-stress true-strain relation and finite-strain approach, an incompressible solution is developed to predict the pressure-deformation relations up to the failure of thick-walled cylinders subjected to internal pressure. The proposed numerical computation procedure is simple and very effective. The total computation times (execution time) for 12 thick-walled cylinders (ρ = 1.25, 1.5, ..., 3.75, 4.00) made of SAE 1045 steel and OFHC copper are 2.871 and 14.195 CP seconds, respectively.

The bursting pressures and pressure-expansion curves for thick-walled cylinders with various radius ratios were presented graphically. This information can readily be used by the designer.

REFERENCES

1. Chu, S.C., "A More Rational Approach to the Problem of Elastoplastic Thick-Walled Cylinders," J. Franklin Ins., Vol. 294, No. 1, pp 57-65, 1972.

2. Hannon, B.M. and Sidebottom, O.M., Plastic Behavior of Open-End and Closed-End Thick-Walled Cylinders," ASME Paper 67 WA/PVP-8, 1967.

3. Hill, R., Lee, E.H., and Tupper, S.J., "Plastic Flow in a Closed-End Tube with Internal Pressure," Proc. First U.S. Nat. Cong. Applied Mechanics, 1951.

4. Steele, M.C., "Partially Plastic Thick-Walled Cylinder Theory," J. Appl. Mech., Vol. 74, pp 133-140, 1952.

5. Allen, D.N. and Sopwith, D.G., "The Stresses and Strains in a Partly-Plastic Thick Tube Under Internal Pressure and End Load," Proc. R. Soc. Lond., Series A, Vol. 205, p 69, 1951.

6. Cook, G., "The Stresses in Thick-Walled Cylinders of Mild Steel Over-Strained by Internal Pressure," Proc. Inst. Mech. Engrs, Vol 126, p 107, 1934.

7. Manning, W.R.D. and Chem, A.M.I., "The Overstrain of Tubes by Internal Pressure," Engineering, Vol. 159, pp 101-102 and 183-184, 1945.

8. Crossland, B. and Bones, J.A., "Behavior of Thick-Walled Steel Cylinders Subjected to Internal Pressure," Proc. Inst. Mech. Eng., Vol 172, pp 777-804, 1958.

9. Hannon, B.M. and Sidebottom, O.M., "Plastic Behavior of Open-End and Closed-End Thick-Walled Cylinders," ASME Paper 67-WA/PVP-8, 1967.

10. Faupel, J.H. and Furbeck, A.R., "Influence of Residual Stress on Behavior of Thick-Wall Closed-End Cylinders," Trans. Amer. Soc. Mech Engrs, Vol 75, pp 345-354, 1953.

11. Iterson, F.K. Th. Van, "Plasticity in Engineering," Blackie, London, 1947.

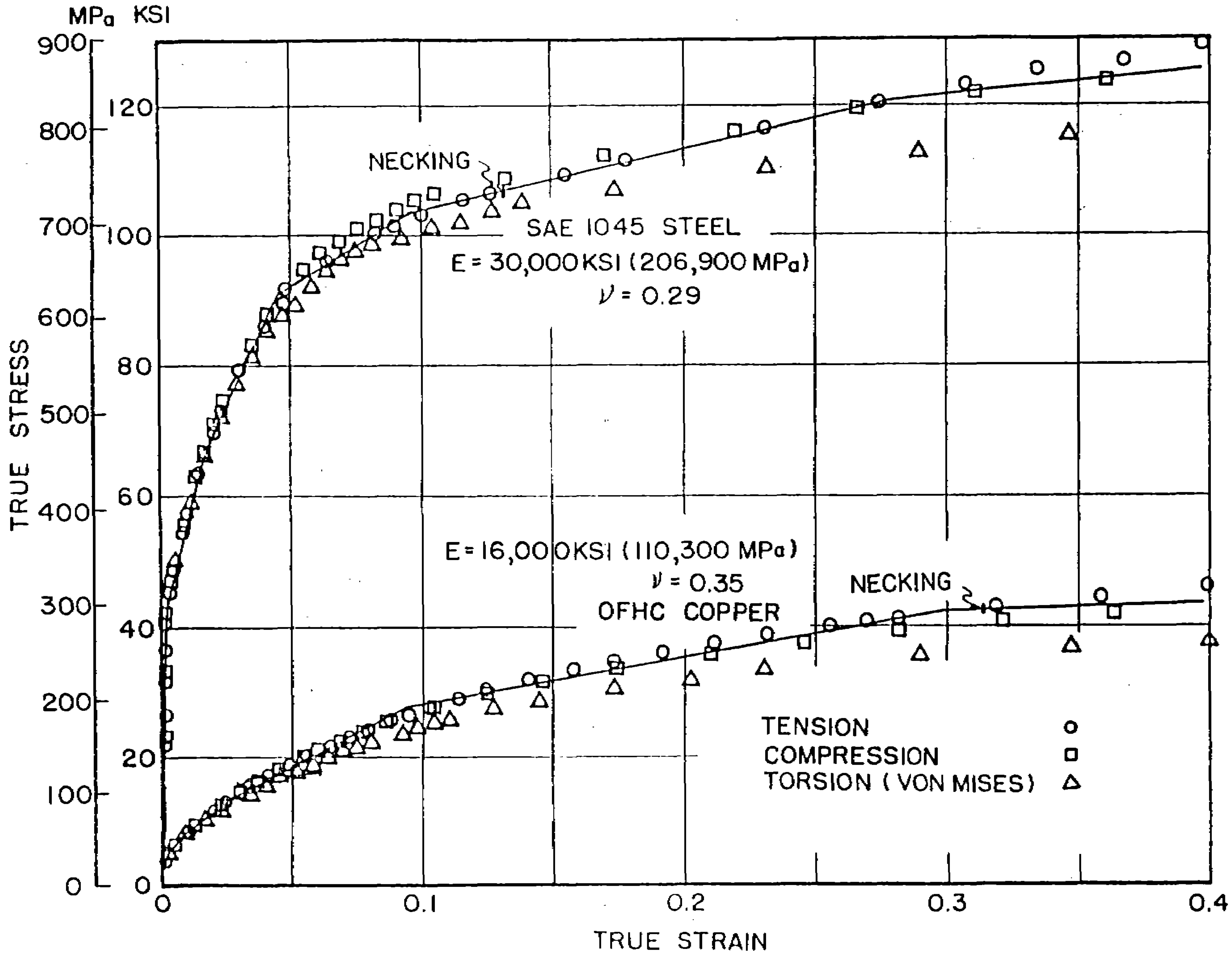


Fig. 1. True-stress vs. true-strain diagrams for SAE 1045 steel and OFHC copper

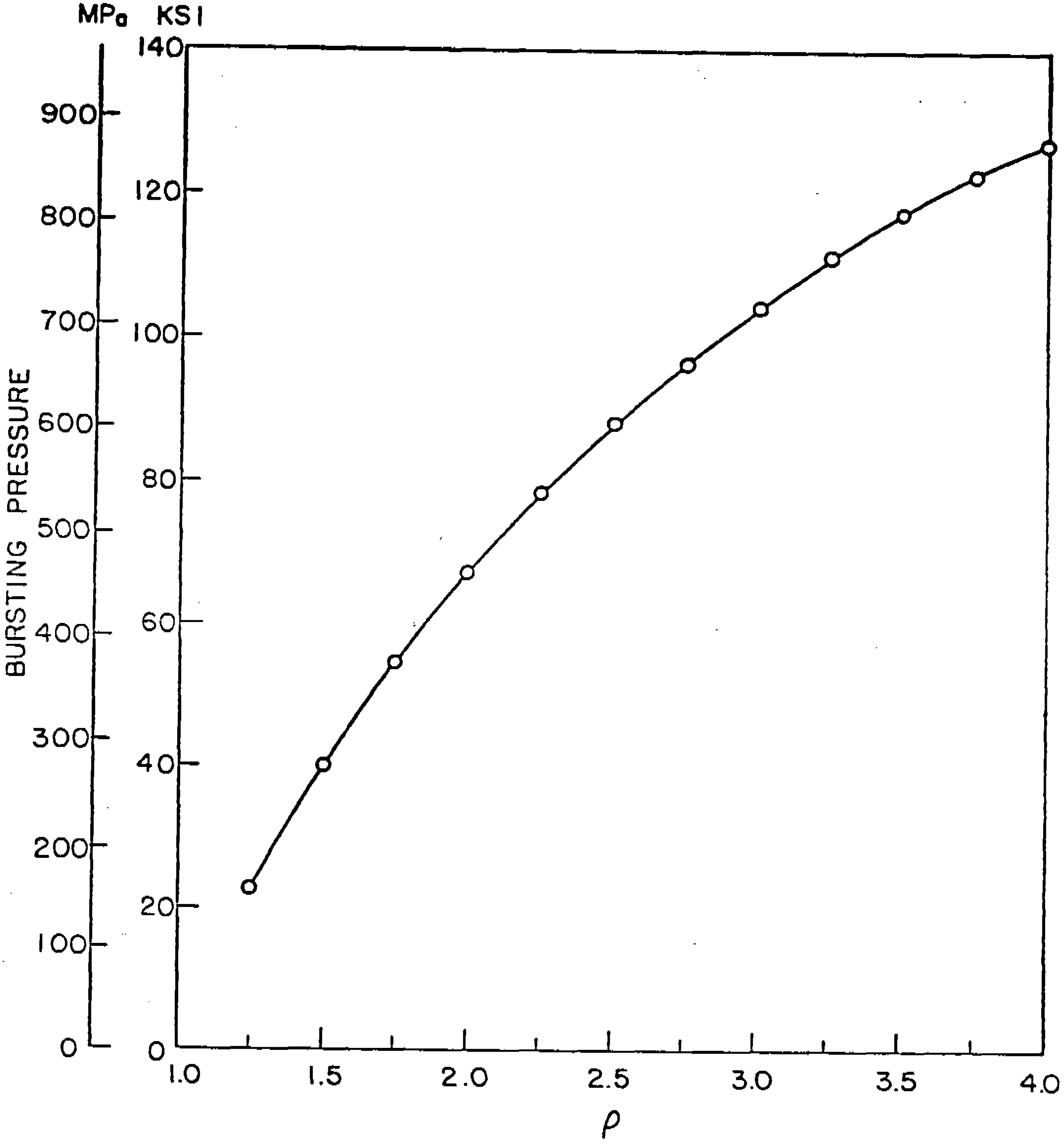


Fig. 2. Bursting pressure of thick-walled cylinders made of SAE 1045 steel

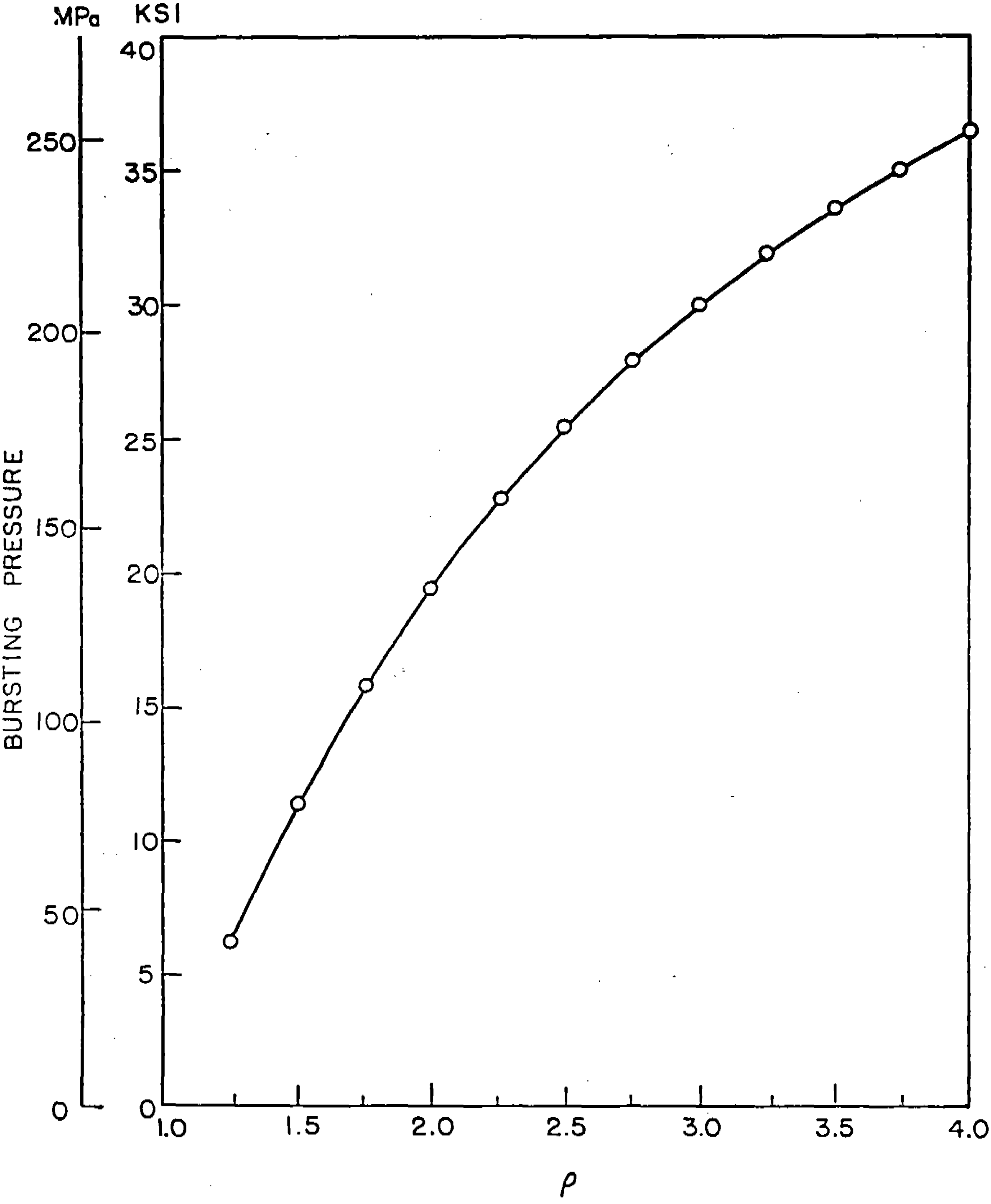


Fig. 3. Bursting pressure of thick-walled cylinders made of OFHC copper

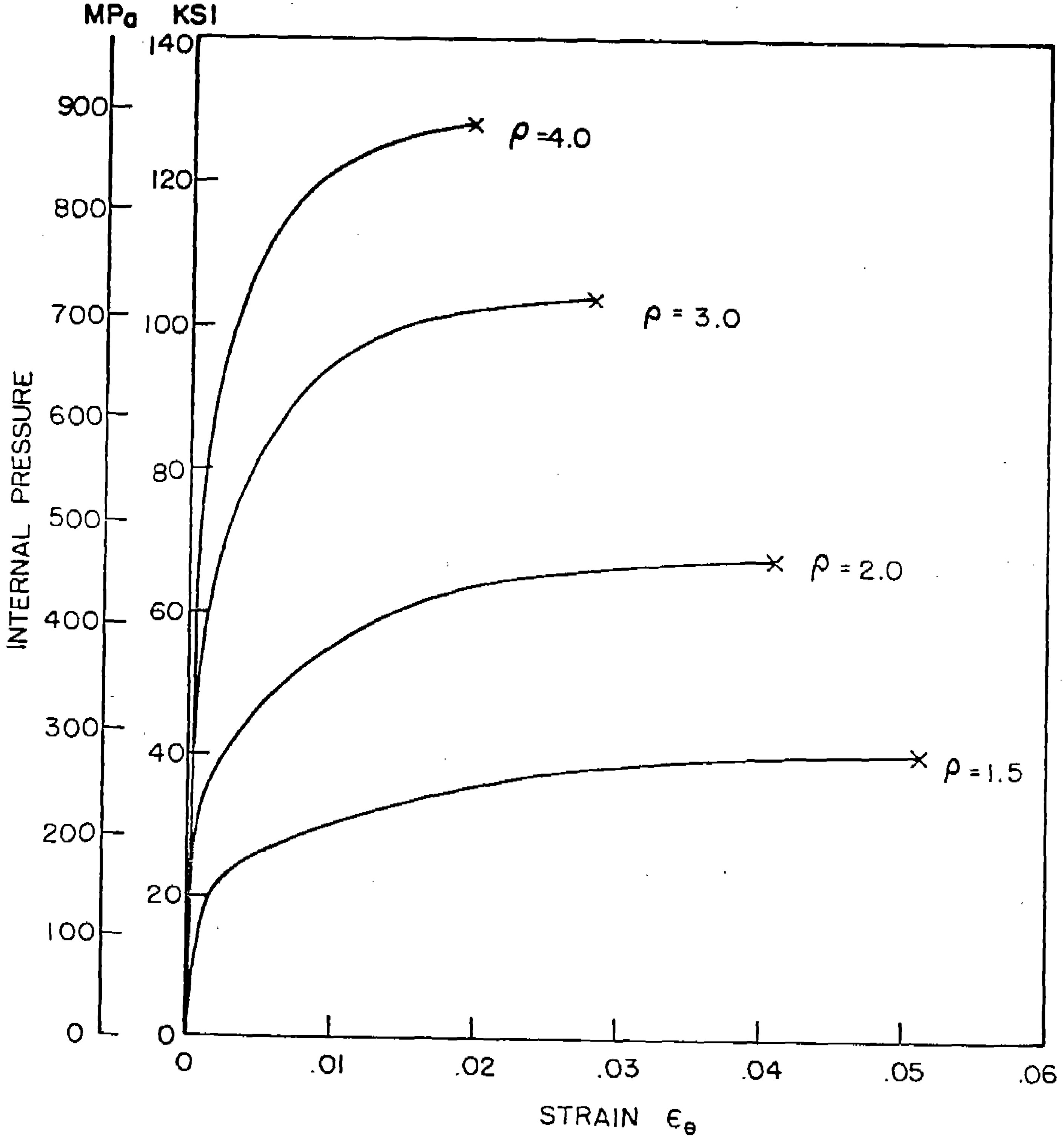


Fig. 4. Pressure-expansion curves of cylinders made of SAE 1045 steel of outer surface

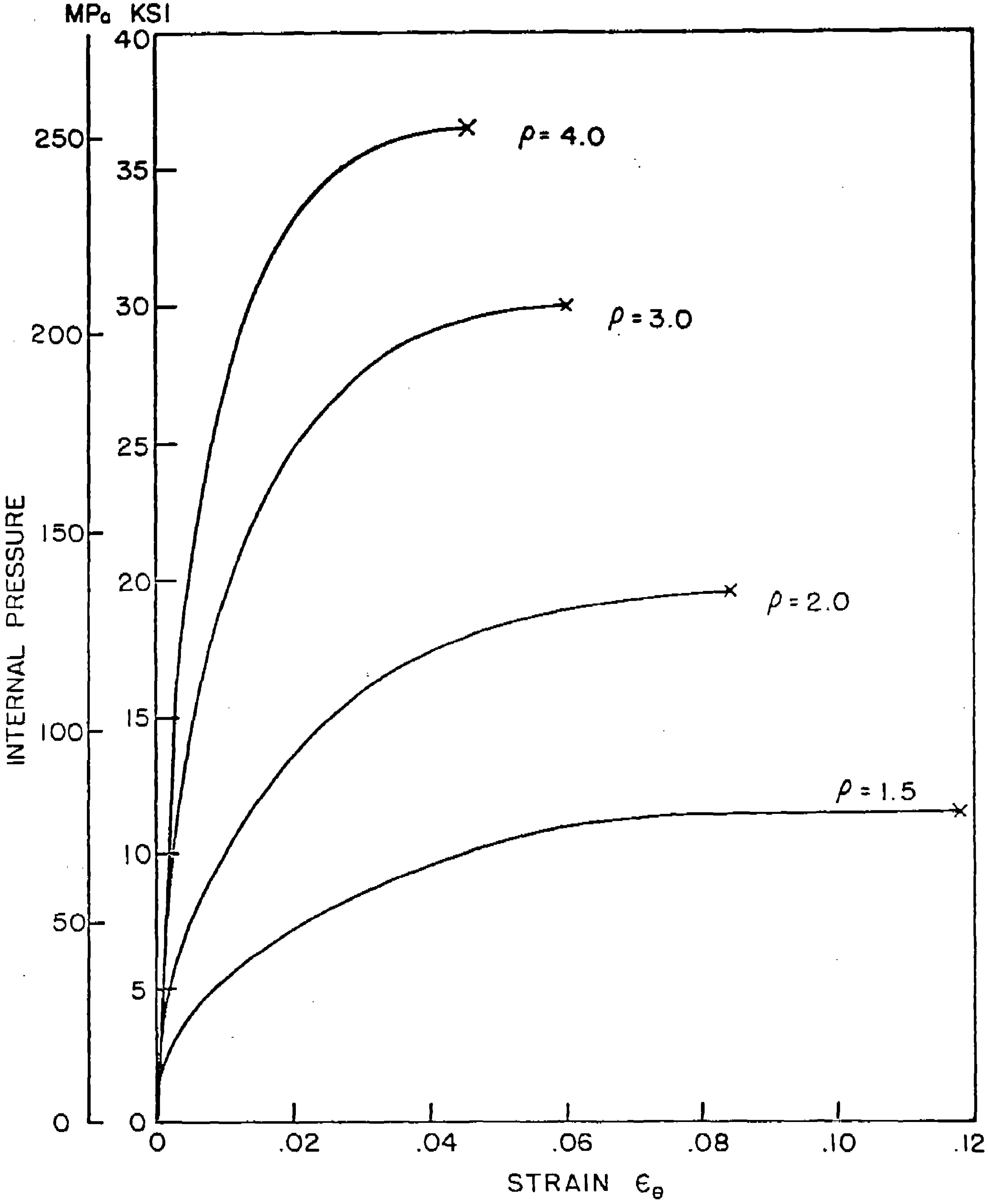


Fig. 5. Pressure-expansion curves of cylinders made of OFHC copper at outer surface

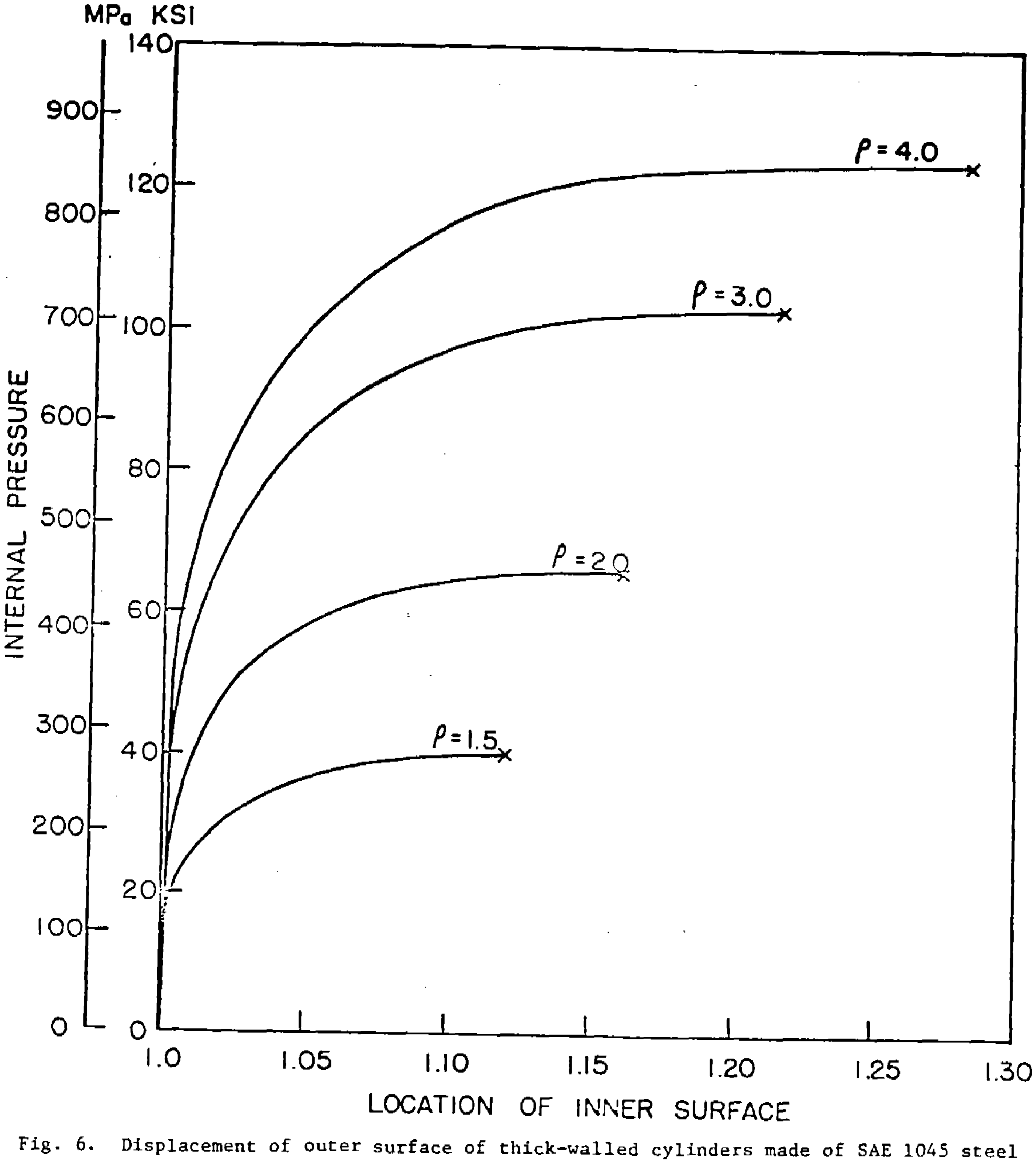


Fig. 6. Displacement of outer surface of thick-walled cylinders made of SAE 1045 steel

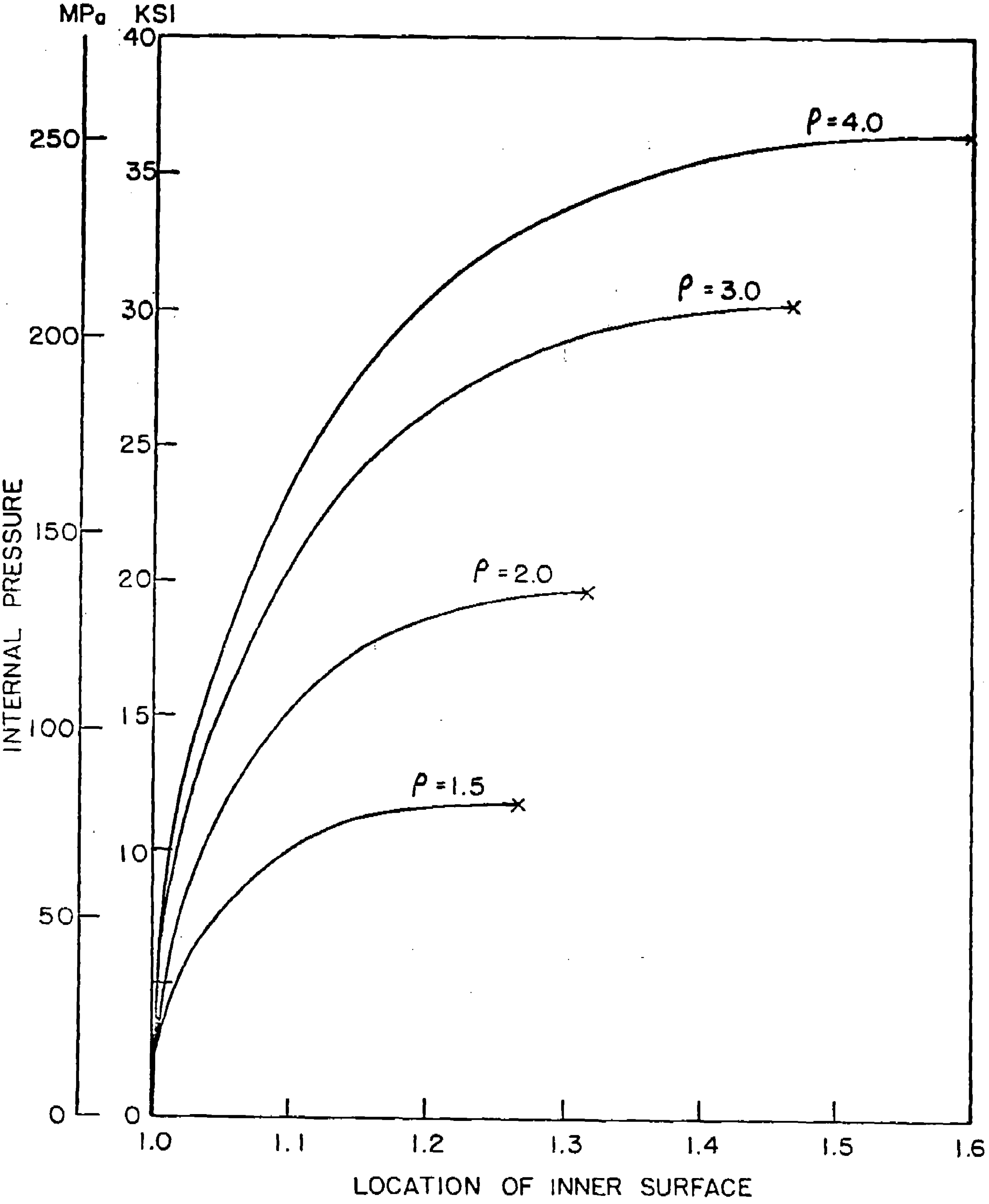


Fig. 7. Displacement of outer surface of thick-walled cylinders made of OFHC copper

STRESS DISTRIBUTION IN A CYLINDRICAL BAR SUBJECTED TO CYCLIC TORSIONAL LOADING

P. C. T. Chen
US Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

H. C. Wu
Division of Materials Engineering
University of Iowa
Iowa City, IA 52242

ABSTRACT. The modified version of the endochronic theory of plasticity is applied to the problem of a cylindrical bar subjected to cyclic fully-reversed torsional loading. The governing equations of integral form are derived for pure shear deformation. Analytical techniques are employed in the solution of these equations. A material with appreciable cyclic hardening behavior is studied. The shear stress-strain curve of such a material under a strain-controlled condition is presented for all cycles until a steady loop is reached. The stress distributions in a cylindrical bar at different stages of loading and unloading are calculated. Some numerical results are presented.

I. INTRODUCTION. In recent years much research has been devoted to developing realistic constitutive equations to describe complex material behavior such as cyclic plasticity [1-9]. The majority of these works are along the lines of the classical theory of plasticity, however, some attempts have also been made in developing new and independent theories. The endochronic theory developed by Valanis [8] is based on the notion of intrinsic time and thermodynamic theory of internal variables. The original definition of intrinsic time has led to difficulties in cases where the history of deformation involves unloading. Valanis [9] has since introduced a new concept of intrinsic time to overcome these difficulties. The new theory has been successfully applied to describe the cyclic hardening phenomenon under uniaxial loading [10].

In this paper the modified version of the endochronic theory of plasticity is applied to the problem of a cylindrical bar subjected to cyclic fully-reversed torsional loading. The governing equations of integral form are presented for pure shear deformation. Analytical techniques are employed in the solution of these equations. The stress distributions in a cylindrical bar at different stages of loading, unloading, and reloading are calculated. The solution for this problem based on a numerical iterative technique was reported recently [11]. A different approach and additional numerical results are to be presented here.

II. ENDOCHRONIC THEORY. According to the modified version of the endochronic theory of plasticity [9], the intrinsic time for the one-dimensional case is defined by

$$d\zeta = \left|d\gamma - k_1 \frac{d\tau}{\mu_o}\right| \tag{1}$$

where k_1 is a positive scalar such that $0 \leq k_1 \leq 1$, τ and γ are shear stress and strain respectively, and μ_o is the shear modulus. When $k_1 = 0$, the original form of intrinsic time is recovered; but when $k_1 = 1$, $\Omega = \gamma - (\tau/\mu_o)$ is the plastic strain, and

$$d\zeta = |d\Omega| \tag{2}$$

Equation (2) is used throughout this investigation so that the concept of plastic strain may be incorporated into the present development.

In the case of pure shear, the governing constitutive equation of integral form is given by

$$\tau = \mu_o \int_o^z \rho(z-z') \frac{d\Omega}{dz'} dz' \tag{3}$$

where

$$\mu_o\rho(z) = \mu_o\rho_o \; \delta(z) + \mu_1 e^{-\alpha z} + \mu_2 \tag{4}$$

in which ρ_o, α, μ_1, and μ_2 are material parameters, $\delta(z)$ is the delta function.

A more general form of the intrinsic time measure involving the strain rate effect is proposed in [10] as

$$d\zeta = k(|\dot{\Omega}|) \; |d\Omega| \tag{5}$$

where k, the strain rate sensitivity function, is a function of $\dot{\Omega}$. In this case, Eq. (3) can be written as

$$\tau = \mu_o \int_o^z \rho(z-z')[\pm \frac{1}{k}] \frac{d\zeta}{dz'} dz' \tag{6}$$

The constitutive Eq. (6) is applied in this paper to describe the material behavior at constant plastic strain rate. The intrinsic time z is related to ζ by the following time scale

$$\frac{d\zeta}{dz} = f(\zeta) \tag{7}$$

where $f(\zeta)$ describes isotropic hardening and is, therefore, termed the hardening function. In this paper, the form

$$f(z) = c - (c-1)e^{-\beta z} \tag{8}$$

will be used because of its simplicity and its proved usefulness in cases of cyclic loading [10]. The parameters c and β are material constants.

III. CYCLIC SHEAR RESPONSE. By using the hardening function given in Eq. (8), the constitutive equation for loading with constant plastic strain rate is written as

$$\tau = (\mu_o/k)\int_o^z \rho(z-z')f(z')dz' \tag{9}$$

where $\rho(z)$ is defined by Eq. (4) and k is a constant. Integrating Eq. (9), the following explicit result is obtained:

$$\tau = (\tau_y/k)f(z) + (\mu_1/k)(g(z)-g(o)) + (\mu_2/k)h(z) \tag{10}$$

where

$$\tau_y = \mu_o\rho_o \text{ is the yield stress} \tag{11}$$

$$h(z) = cz + \beta^{-1}(c-1)(e^{-\beta z}-1) \tag{12}$$

$g(z)$, $g(o)$ are the values of the function $g(z')$ evaluated at $z' = z, o$, respectively, and

$$g(z') = (c/\alpha)e^{-\alpha(z-z')} - (c-1)/(\alpha-\beta)e^{-\alpha z+(\alpha-\beta)z'} \tag{13}$$

Equation (10) is now the response function for loading. If unloading occurs when the plastic strain reaches Ω_1 (or $z = z_1$), then the response function for $z > z_1$ can be obtained as

$$\tau = -(\tau_y/k)f(z) + (\mu_1/k)[-g(z) + 2g(z_1) - g(o)] + (\mu_2/k)[-h(z) + 2h(z_1)] \tag{14}$$

If Eqs. (10) and (14) are examined at $z = z_1^-$ and z_1^+, respectively, a drop in stress of $2(\tau_y/k)f(z_1)$ results during elastic unloading.

If reloading takes place after unloading and, assuming that reloading occurs at $\Omega = \Omega_2$ (or $z = z_2$, where $z_2 > z_1$), the response function for $z > z_2$ is obtained as

$$\tau = (\tau_y/k)f(z) + (\mu_1/k)[g(z) - 2g(z_2) + 2g(z_1) - g(o)]$$
$$+ (\mu_2/k)[h(z) - 2h(z_2) + 2h(z_1)] \tag{15}$$

When Eqs. (14) and (15) are examined at $z = z_2^-$ and z_2^+, respectively, a jump in stress of $2(\tau_y/k)f(z_2)$ is again obtained during elastic reloading.

This procedure of obtaining a theoretical expression for each part of the cyclic loading process may be continued. Thus, a general expression of the response function for the constant total strain amplitude cyclic torsion test may be found to be

$$\tau = (-1)^N(\tau_y/k)f(z) + (\mu_1/k)G(z) + (\mu_2/k)\Omega(z) \tag{16}$$

where

$$G(z) = (-1)^N[g(z)-g(z_N)] + \sum_{n=1}^{N} (-1)^{n+1}[g(z_n) - g(z_{n-1})] \tag{17}$$

$$\Omega(z) = (-1)^N[h(z) - h(z_N) + \sum_{n=1}^{N} (-1)^{n+1}[h(z_n) - h(z_{n-1}) \tag{18}$$

Note that $N \neq 0$ is the number of half-cycles, odd for unloading and even for reloading.

After many cycles when the values of z become large, the hysteresis loop of the stress-strain curve will approach a steady state. If Eq. (16) is examined at $z = z_N^-$ and $z = z_N^+$, a drop or jump in stress of magnitude $2(\tau_y/k)f(z_N)$ results, which corresponds to the elastic response upon the reversal of the loading or unloading direction. When z_N is sufficiently large, the jump or drop of stress becomes a constant value $2c\ \tau_y/k$ at the steady state.

IV. CYLINDRICAL BAR UNDER TORSION. For a cylindrical bar under torsional loading, the external torque is

$$T_s = 2\pi \int_0^{r_a} \tau r^2 dr \tag{19}$$

where τ is the current shear stress corresponding to location r, and r_a is the radius of cross-section. Geometric considerations show that radial lines have to remain straight after deformation. Thus, one concludes that

$$\gamma = (\gamma_a/r_a)r \tag{20}$$

where γ_a is the strain at the outermost fiber. Since a yield stress is introduced in Eq. (4), an elastic core always exists during deformation whose radius r_e is given by

$$r_e = \frac{\tau_y}{\mu_o}\frac{r_a}{\gamma_a} \tag{21}$$

If the experiment is strain-controlled with strain at r_a varying between $-\gamma_a$ and $+\gamma_a$, and with γ_a in the plastic range, then the torque can be computed as

$$T_s = \frac{\pi}{2}\tau_y r_e{}^3 + 2\pi(r_a/\gamma_a)^3 \int_{\gamma_e}^{\gamma_a} \tau\gamma^2 d\gamma \tag{22}$$

where

$$\gamma_e = (\gamma_a/r_a)r_e \quad , \quad \gamma = \Omega + \tau/\mu_o \tag{23}$$

τ and Ω are given by Eqs. (16) and (18), respectively.

Now that γ is known, the values of z and τ at each fiber can be calculated and used in Eq. (22). Note that there exists an explicit expression for all cycles

$$d\gamma/dz = \Omega'(z) + \frac{1}{\mu_o}\tau'(z) = p(z) \tag{24}$$

where

$$\Omega'(z) = (-1)^N f(z) \quad , \quad f'(z) = \beta(c-1)e^{-\beta z}$$

$$\tau'(z) = (-1)^N[\tau_y/k)f'(z) + (\mu_2/k)f(z)] + (\mu_1/k)\Gamma'(z)$$

$$\Gamma'(z) = (-1)^N[F'_{zz} - F'_{Nz}] - \sum_{n=1}^{N}(-1)^n[F'_{nz} - F'_{(n-1)z}]$$

$$F'_{zz} = [\beta(c-1)/(\alpha-\beta)]e^{-\beta z} \quad , \quad F'_{nz} = -\alpha F_{nz}$$

and

$$F_{nz} = \frac{c}{\alpha}e^{-\alpha(z-z_n)} - (\frac{c-1}{\alpha-\beta})e^{-\alpha z + (\alpha-\beta)z_n} \tag{25}$$

The integral in Eq. (22) can be replaced by

$$\int_{\gamma_e}^{\gamma_a}\tau\gamma^2 d\gamma = \int_{o}^{z_a}\tau\gamma^2 p(z)dz \tag{26}$$

where z_a can be calculated by Eq. (23) with known value γ_a. Now the numerical integration becomes very easy because the values of the integrand can be evaluated directly.

V. NUMERICAL RESULTS AND DISCUSSION. To apply the developed model the material constants (α, β, c, τ_y, μ_o, μ_1, μ_2) in the theory have to be determined. These material constants can be determined if the cyclic shear stress-strain curve for the material has been obtained experimentally. The usual procedure is to perform a cyclic torsion test using a thin-walled tubular specimen. Such a test for the annealed AISI 4142 steel was carried out by the Plasticity Research Laboratory at the University of Iowa. The values of constants were then used to predict the results for a solid bar test. The theoretical and experimental results were in reasonable agreement [11]. This real material does not show any appreciable amount of cyclic hardening.

For purpose of investigating the implications of the developed model, a hypothetical material with appreciable cyclic hardening behavior was studied. The shear stress-strain behavior of such material under fully-reversed torsional loading is presented in Figure 1. The material constants were determined as: $\alpha = 1000$, $\beta = 50$, $c = 1.5$, $\tau_y = 10^4$ psi, $\mu_o = 10^7$ psi, $\mu_1 = 4\text{x}10^6$ psi, $\mu_2 = 0$. A steady loop is established after a few cycles. The same set of constants were then used to predict the stress distribution in a cylindrical bar subjected to cyclic torsional loading. We have carried out the computational process seven cycles after initial loading. The numerical results are presented here for the initial loading half-cycle and two cycles of unloading and reloading.

Figure 2 presents the numerical results for the torque as related to the shear strain at the outermost fiber. The distribution of stress in the cross-section at different magnitudes of torque during the initial loading half-cycle is presented in Figure 3. The corresponding shear strains at the outer fiber are 0.1, 0.3, 0.6, and 1.0 percent, respectively. Notice that the outer fiber is the first one to yield at $\gamma_a = 0.1$ percent; subsequently as more torque is applied, the radius of the elastic inner core gets smaller. Also notice that the rate of hardening for each fiber decreases which is, of course, in accordance with strain hardening phenomena.

Figure 4 presents the distribution of the shear stress in the bar at three stages of the first unloading half-cycle. The top and bottom curves correspond to the beginning and end of the unloading stages, i.e., $\gamma_a = \pm 1$ percent. The corresponding values of torque are 1256, -1387 lb-in., respectively. The middle curve represents the residual stress distribution when the applied torque is equal to zero. A few iterations are needed to reach this state and the residual shear strain at the outermost fiber is 0.378 percent. Figure 5 presents the distribution of the shear stress in the bar at three stages of the first reloading half-cycle, i.e., $T_s = -1387$, 0, 1461 lb-in. The middle curve represents the distribution of residual stress when the torque is equal to zero. Reverse yielding occurs within the outer 12 percent of the section when $T_s = 0$ during unloading and reloading. The distribution of residual stress during unloading is quite different from that during reloading as shown in Figures 4 and 5 for the first cycle. Similar results for the residual stress distribution during the second unloading-reloading cycle are shown in Figure 6. The solid curve is the favorable one if the applied torque during service is in the same direction as the initial loading.

REFERENCES

1. Z. Mroz, "Simplified Theories of Cyclic Plasticity," Acta Mechanica, Vol. 22, pp. 131-152, 1975.

2. A. Miller, "An Inelastic Constitutive Model for Monotonic, Cyclic, and Creep Deformation: Part I - Equations, Development, and Analytical Problems," J. Eng. Materials and Tech., Vol. 98, pp. 97-105, 1976.

3. M. A. Eisenberg, "A Generalization of Plastic Flow Theory With Application to Cyclic Hardening and Softening Phenomena," J. Eng. Materials and Tech., Vol. 98, pp. 221-228, 1976.

4. E. Krempl, M. C. M. Liu, and D. C. Nairn, "An Exponential Stress-Strain Law for Cyclic Plasticity," J. Eng. Materials and Tech., Vol. 98, pp. 322-329, 1976.

5. E. P. Popov and H. Peterson, "Cyclic Metal Plasticity: Experiment and Theory," J. Eng. Mech. Div., Proc. ASCE, Vol. 104, EM6, pp. 1371-1388, 1978.

6. S. R. Bodner and I. Partom, "Uniaxial Cyclic Loading of Elasto-Viscoplastic Materials," J. Appl. Mech., Vol. 46, pp. 805-810, 1979.

7. D. C. Drucker and L. Palgen, "On Stress-Strain Relations Suitable for Cyclic and Other Loading," J. Appl. Mech., Vol. 48, pp. 479-485, 1981.

8. K. C. Valanis, "A Theory of Viscoplasticity Without a Yield Surface, Part I and Part II," Archives of Mechanics, Vol. 23, pp. 517-551, 1971.

9. K. C. Valanis, "Fundamental Consequences of a New Intrinsic Time Measure-Plasticity as a Limit of the Endochronic Theory," Archives of Mechanics, Vol. 32, pp. 171-191, 1980.

10. H. C. Wu and M. C. Yip, "Endochronic Description of Cyclic Hardening Behavior for Metallic Materials," J. of Eng. Materials and Technology, Vol. 103, pp. 212-217, 1981.

11. P. C. T. Chen, M. R. Aboutorabi, and H. C. Wu, "Cyclic Torsion of a Circular Cylinder," Proc. of 8th Army Symposium on Solid Mechanics, AMMRC MS82-4, pp. 405-415, 1982.

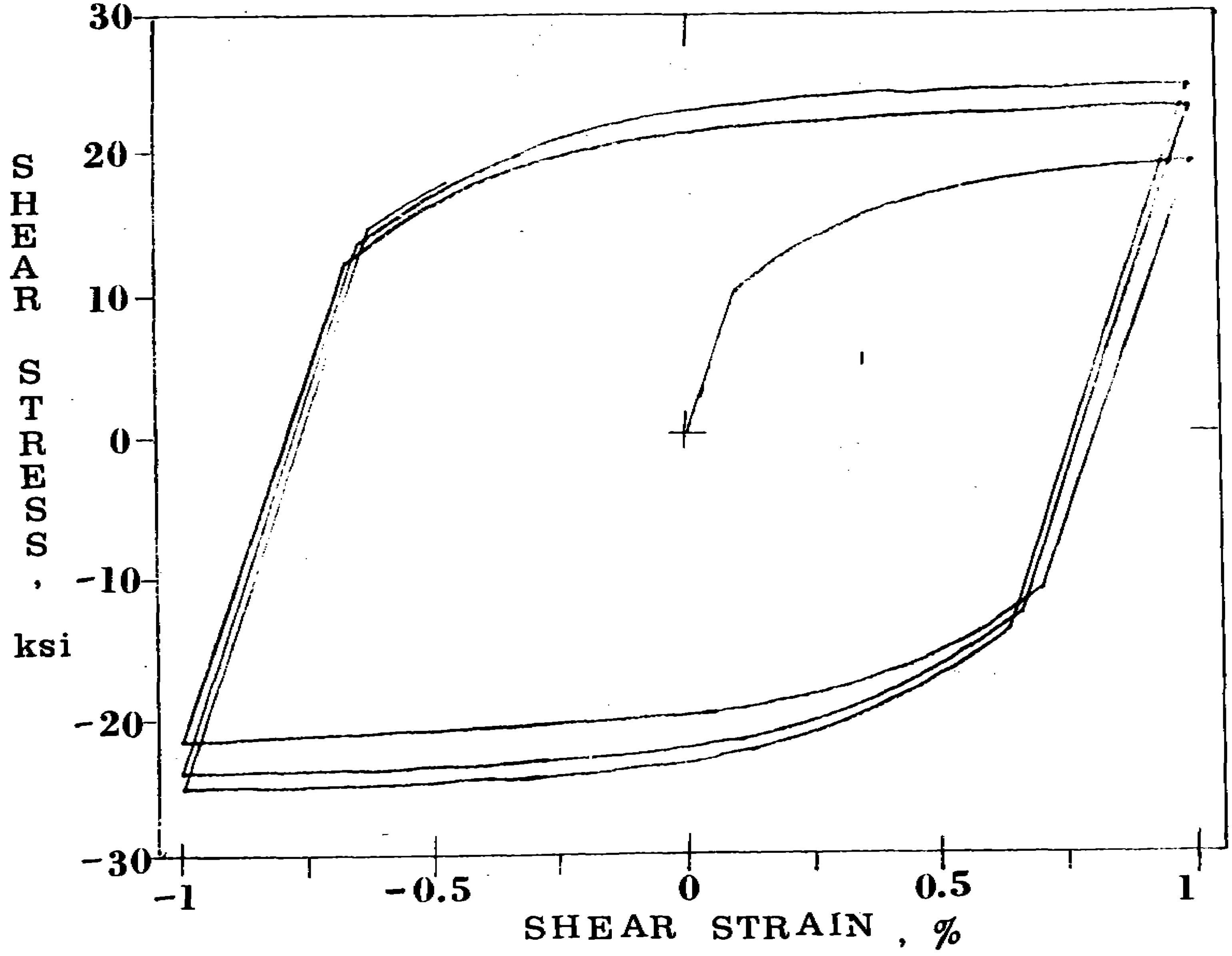


Figure 1. Cyclic Shear Stress-Shear Strain Curve.

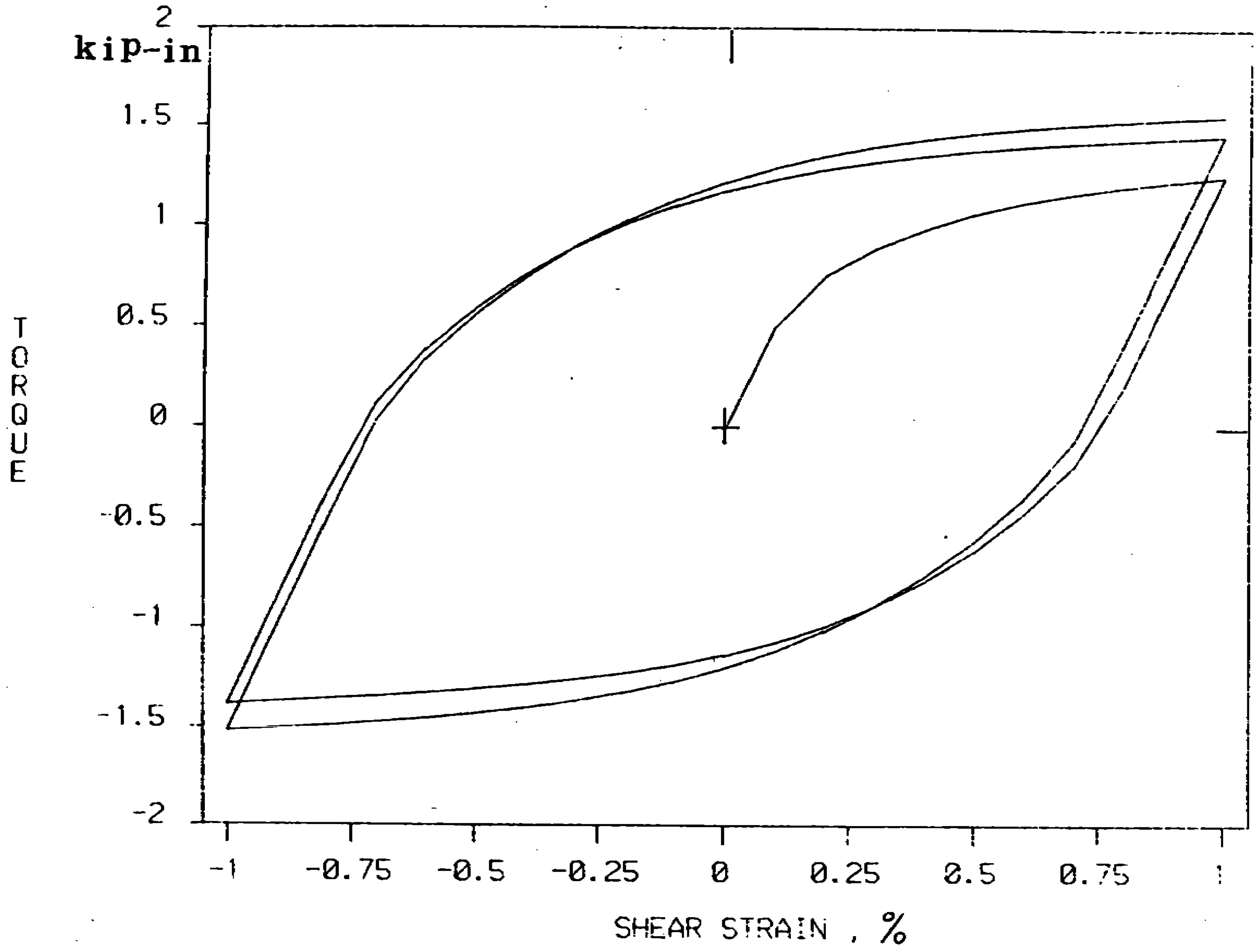


Figure 2. Torque vs. Outer Shear-Strain Curve.

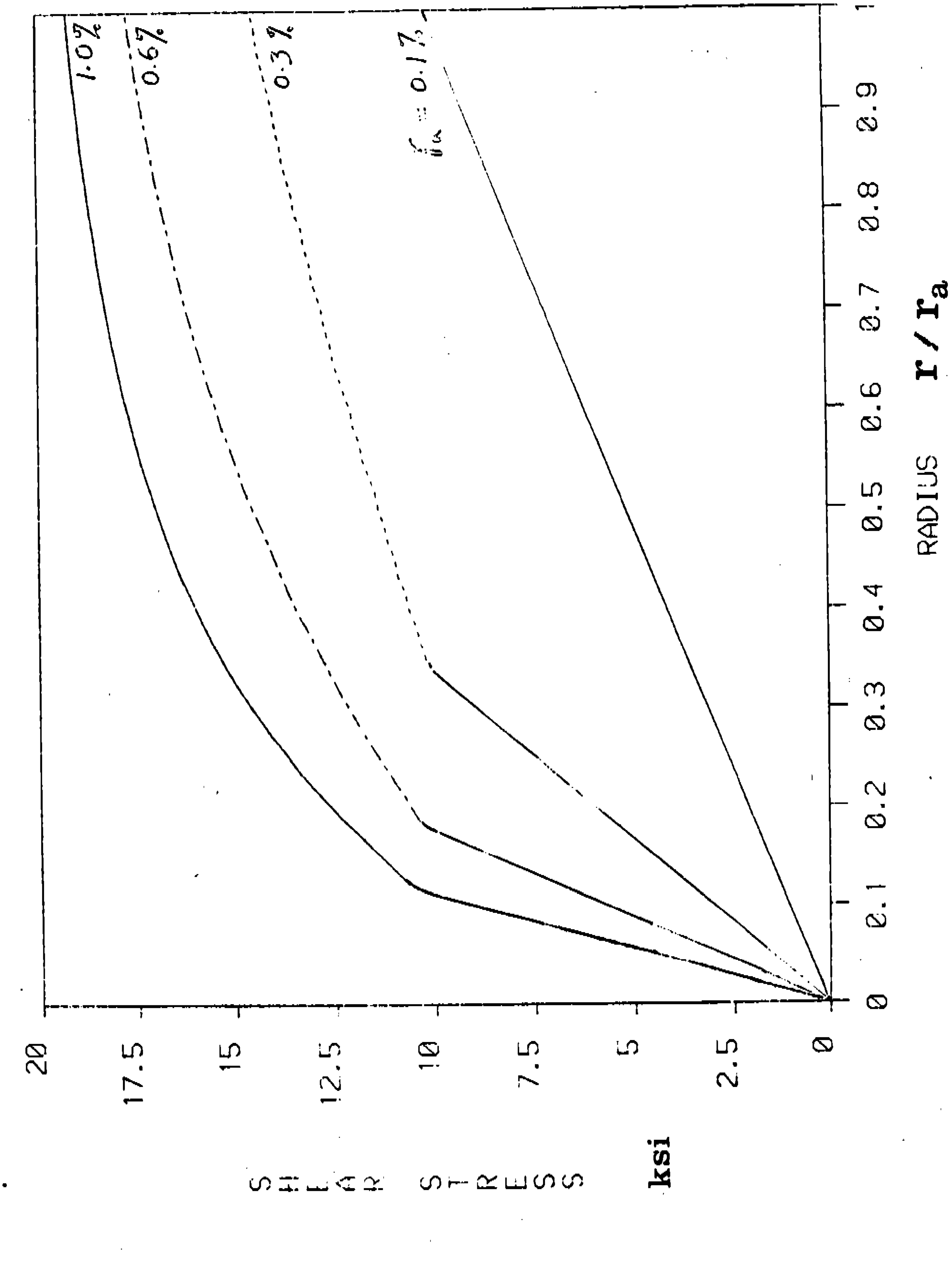


Figure 3. Stress Distribution During Initial Loading.

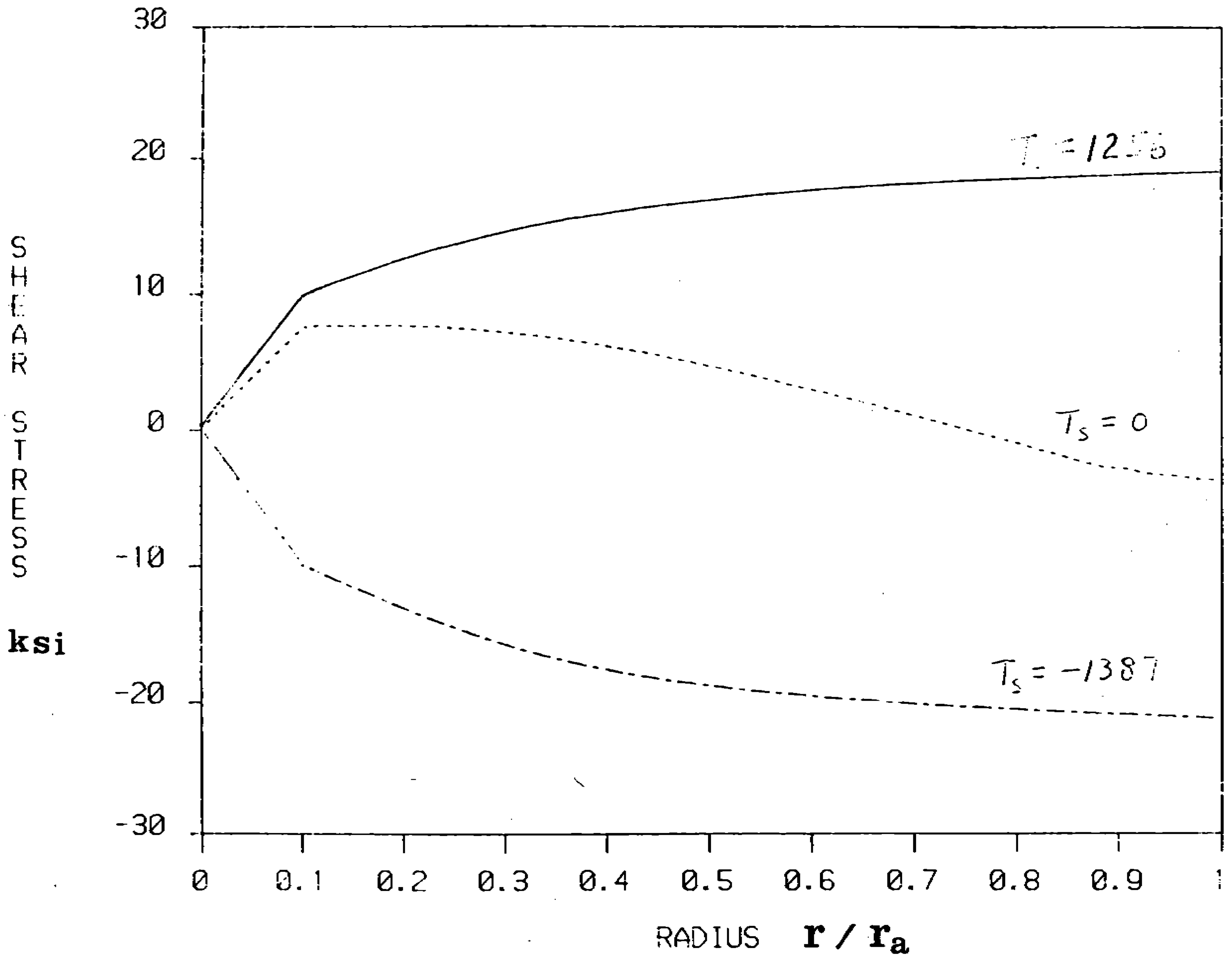


Figure 4. Stress Distribution During First Unloading.

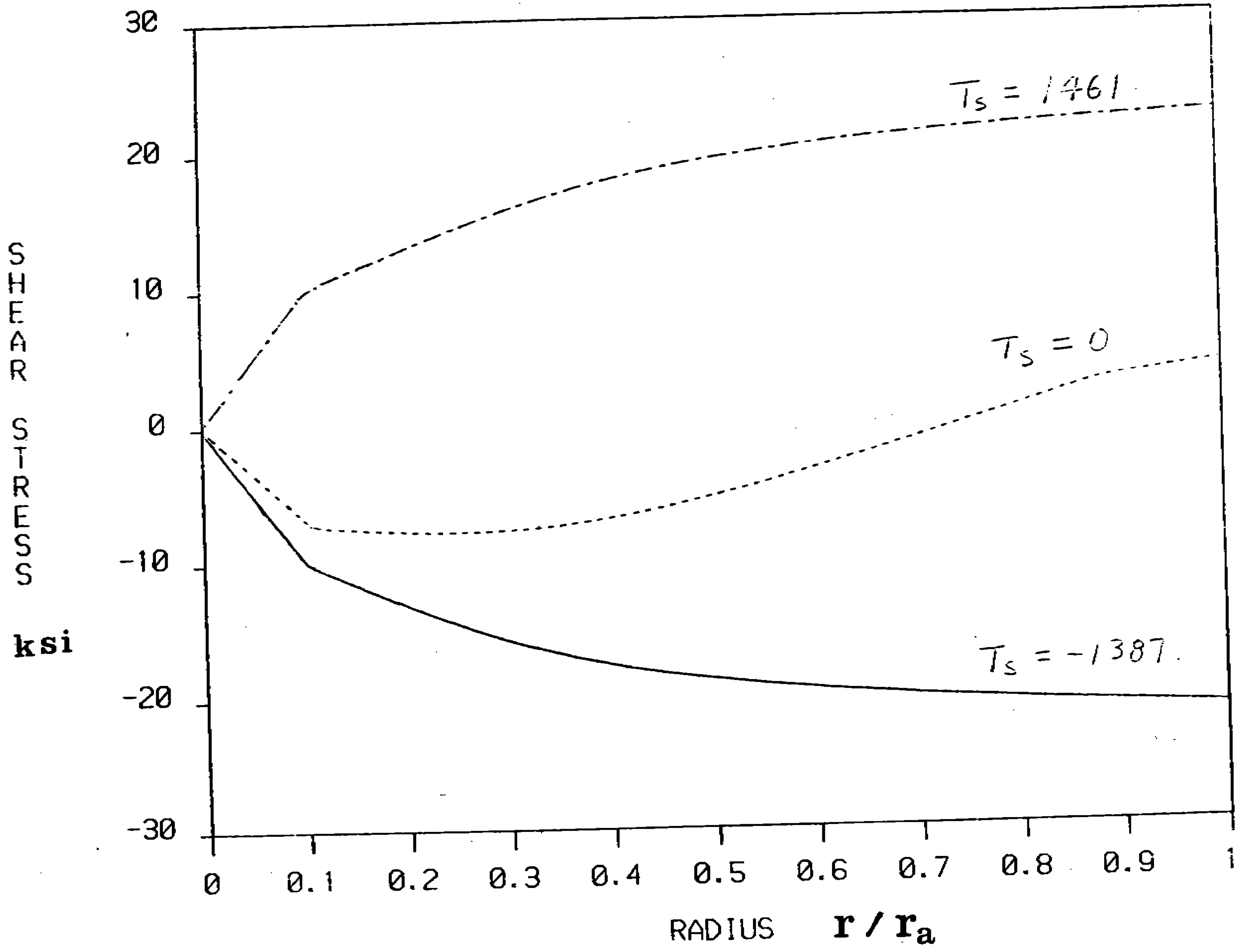


Figure 5. Stress Distribution During First Reloading.

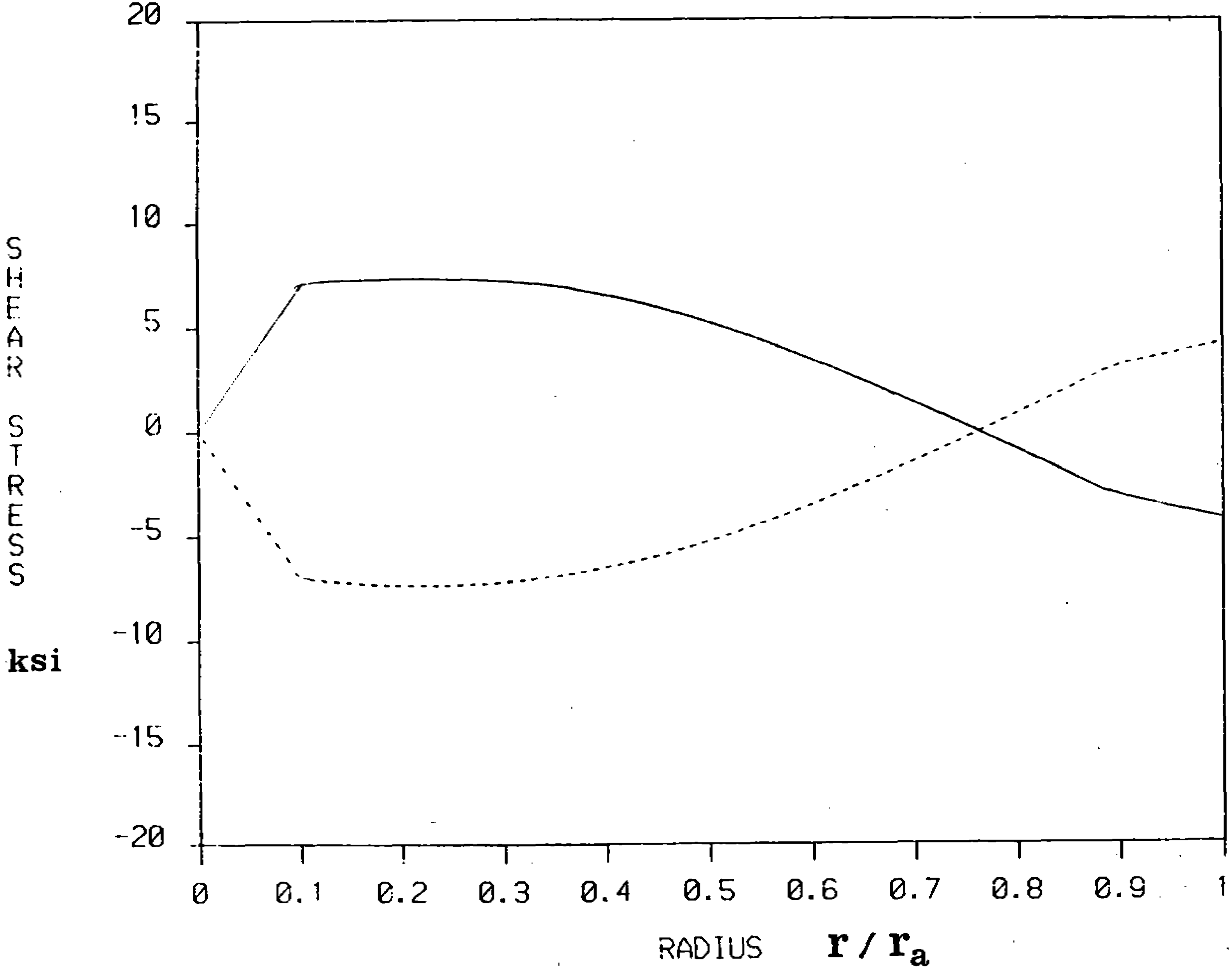


Figure 6. Distribution of Residual Stresses in Second Cycle.

FINITE ELEMENT RESULTS OF PRESSURIZED THICK TUBES BASED ON TWO ELASTIC-PLASTIC MATERIAL MODELS

P. C. T. Chen and G. P. O'Hara
US Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. The loading and unloading problems in thick tubes subjected to uniform internal pressure have been analyzed with the ADINA finite element code. The elastic-plastic materials are modeled by two strain-hardening rules - isotropic and kinematic. The von Mises yield condition, the associated flow theory, and a multi-linear stress-strain curve are used in both material models. The numerical results of the stresses and displacements for thick tubes with different wall ratios are obtained as functions of loading history. A comparison of numerical results based on two material models is made.

I. INTRODUCTION. The problem of pressurized thick-walled tubes is of practical importance to pressure vessels and the autofrettage process of gun barrels. Many solutions for this problem have been reported over the last three decades [1-7]. This is a result of different mathematical methods, end conditions, and material models. Different assumptions for the material properties such as compressibility, yield criterion, flow rule, hardening rule, etc. can lead to many material models. A common feature in all earlier investigations is to introduce certain restrictive assumptions so as to simplify the mathematical analysis [1-4]. The recent development in numerical methods makes it possible to use more realistic material model and to consider more general problems. Both the finite element method [5] and the finite difference method [6,7] have been used to solve the elasto-plastic problems with different end conditions and more general loading conditions. The material model was based on the von Mises yield criterion, the Prandtl-Reuss flow theory, and the isotropic hardening rule.

The finite element method is more powerful and can be used to solve more general nonlinear problems [8,9]. Many finite element codes have been developed as seen in a recent survey paper [10]. The ADINA code, developed by K. J. Bathe, is a general purpose finite element program for Automatic Dynamic Incremental Nonlinear Analysis [11]. The standard version models the elastic-plastic behavior of metals by the use of the Mises yield criterion, the associated flow theory, and two strain-hardening rules - isotropic and kinematic. Both hardening models were limited to linear hardening in our first version acquired in 1981. The multi-linear option was allowed in our second version one year later. This paper shows an application of the ADINA code to our pressurized thick tube problems. A multi-linear stress-strain curve is used in both material models and thick tubes of different wall ratios are considered. The numerical results together with a brief summary of the elastic-plastic theory, finite element formulation are presented below with emphasis on the basic assumptions used. More detailed theoretical information can be found in a forthcoming report [12].

II. ELASTIC-PLASTIC THEORY. In elastic-plastic analysis the material behavior is described using three properties in addition to the elastic stress-strain relations, namely a yield criterion, a flow rule, and a hardening rule.

The initial and subsequent yield condition for isothermal kinematic or isotropic hardening can be written as

$$f(\sigma_{ij}-\alpha_{ij}) - \sigma(\int d\varepsilon^{\rho}) = 0 \qquad (1)$$

where σ_{ij} is the stress tensor, α_{ij} is a tensor denoting the translation of the yield surface, f is the yield function, and $\sigma(\int d\varepsilon^{\rho})$ represents the dependence of the yield stress on the accumulated increments of effective plastic-strain. The von Mises yield function for kinematic hardening is

$$f = [\frac{3}{2} (s_{ij}-\alpha_{ij})(s_{ij}-\alpha_{ij})]^{1/2} \qquad (2)$$

where

$$s_{ij} = \sigma_{ij} - \frac{1}{3} \sigma_{kk}\delta_{ij}$$

and

$$\alpha_{ij} = 0 \quad \text{for isotropic hardening} \qquad (3)$$

Restricting the analysis to associated flow rules, the plastic strain increment $d\varepsilon_{ij}{}^{\rho}$ is derivable from the plastic potential function f by

$$d\varepsilon_{ij}{}^{\rho} = q_{ij}d\lambda \quad \text{and} \quad q_{ij} = \partial f/\partial\sigma_{ij} \qquad (4)$$

where $d\lambda$ is a scalar to be determined.

During active plastic deformation the yield function must be satisfied continuously, so that the consistency condition is

$$(d\sigma_{ij}-d\alpha_{ij})\partial f/\partial\sigma_{ij} = 0 \qquad (5)$$

The original kinematic hardening concept was Prager's rule [13] that

$$d\alpha_{ij} = (\frac{2}{3} H')d\varepsilon_{ij}{}^{\rho} \quad \text{and} \quad H' = d\sigma/d\varepsilon^{\rho} \qquad (6)$$

Prager's rule was used in the ADINA formulation although its modification by Ziegler [14] is more popular. Equations (1) through (6) are the basic equations of the elastic-plastic theory. In addition, we need the elastic stress-strain relation

$$d\sigma_{ij} = E_{ijmn}(d\varepsilon_{mn}-d\varepsilon_{mn}{}^{\rho}) \qquad (7a)$$

where E_{ijmn} is the elastic constitutive tensor. If the material is initially isotropic, then

$$E_{ijmn} = \frac{E}{1+\nu} [\delta_{im}\delta_{jn} + \frac{\nu}{1-2\nu} \delta_{ij}\delta_{mn}] \tag{7b}$$

where E and ν are the Young's modulus and Poisson's ratio, respectively.

Using the basic equations (1) to (7), we can obtain the incremental stress-strain relation for elastic-plastic material models

$$d\sigma_{ij} = D_{ijmn}\, d\varepsilon_{mn} \tag{8a}$$

where

$$D_{ijmn} = E_{ijmn} - \frac{E_{ijtu}\, q_{tu}\, q_{vw}\, E_{vwmn}}{H' + q_{kl}\, E_{klrs}\, q_{rs}} \tag{8b}$$

This constitutive relation holds for the combined isotropic-kinematic hardening model. For the special cases using Eqs. (2) to (4), we have

$$\text{isotropic hardening:} \quad q_{ij} = 3s_{ij}/(2\sigma) \tag{9a}$$

$$\text{kinematic hardening:} \quad q_{ij} = 3(s_{ij}-\alpha_{ij})/(2\sigma) \tag{9b}$$

III. FINITE ELEMENT FORMULATION. The finite element formulation used in ADINA is very general that large strain dynamic analysis has been considered [11,12]. Since the present problem requires only a small strain static analysis, a very brief summary of the special formulation is presented here. The geometry of the body is discretized by two-dimensional 8-nodes isoparametric elements. The coordinates and displacements are interpolated by the same shape functions N_i, i = 1 to 8, i.e.,

$$x = N_i\bar{x}_i \quad , \quad u = N_i\bar{u}_i \quad , \quad \text{etc.} \tag{10}$$

where $\bar{x}_i$, $\bar{y}_i$, $\bar{u}_i$, $\bar{v}_i$ are the coordinates and displacements at the nodal points. The strain increments in elements can be obtained by differentiation and in matrix notation we have

$$\{\Delta\varepsilon\} = [B]\{\Delta U\} \quad \text{and} \quad [B] = [L][N] \tag{11}$$

where [L] is a linear differential operator and $\{\Delta U\}$ is a vector of all nodal displacement increments in an element.

Once we know [D] and [B], we can compute the element stiffness matrix by

$$[K] = \int_V [B]^T[D][B]\, d(\text{vol}) \tag{12}$$

To carry out numerical integration, we express all matrices and volume element in terms of local coordinates and evaluate them at integration stations with the aid of Gauss quadrature formulae. For double summation we use either

(2x2) or (3x3) points in a rectangle. This finite element formulation is based on displacements so the kinematic equations and constitutive equations are satisfied locally. The principle of virtual displacements is used to express the equilibrium of the body in the current configuration. Since the principle is in integral form, we can sum all element contribution to the system.

IV. THICK TUBES. Consider a long thick tube, internal radius a, and external radius b, which is subjected to internal pressure p. Thick tubes of different wall ratios are considered. The geometry of the tube is discretized by two-dimensional axisymmetric 8-nodes isoparametric elements along the radial direction. We use 10 elements for smaller wall ratios (b/a = 1.5 and 2.0) and 20 elements for larger wall ratios (b/a = 3.0 and 4.63). All elements are of equal size and 3x3 points are used in carrying out the numerical integration. The displacements at the nodal points and the stresses at the integration points are obtained as functions of loading history. At each stage of loading, we have N+1 results for the displacements and 3N results for the stresses where N is the number of elements used.

The common input data for both material models are E = $2.583\text{x}10^7$ psi, ν = 0.3, and 6 points on the uniaxial stress-strain curve, i.e., (σ in Ksi, ε in %) = (155, 0.6), (167, 0.85), (172, 1.25), (177, 3.0), (181, 5), (181, 15). These six points are chosen to give a piecewise linear representation to the actual stress-strain curve for a high strength steel as shown in Figure 1. The ADINA code allows a maximum of 7 points to represent two multi-linear hardening models (model number 8 and 9 for isotropic and kinematic hardening). These two hardening models are widely used because of their simplicity. Isotropic hardening is generally considered to be a suitable model for large plastic flows. Kinematic hardening is the simplest theory that can model the Bauschinger effect. If unloading does not occur, there is no difference between these two models. For unloading with reverse yielding, the finite element results based on these two models will be different.

The loading and unloading problems in thick tubes of different wall ratios have been analyzed using the ADINA code and two hardening models. The tubes of wall ratios 1.5 and 2.0 have been loaded to reach fully plastic state and then unloaded completely. No reverse yielding occurs during unloading for tubes with both wall ratios and the usual assumption of elastic unloading is justified on the basis of these two material models. The numerical results for the tube with b/a = 2 are shown in Figures 2 through 4. Figure 2 shows the boundary displacements (U_a and U_b) as functions of pressure history. We use 11 steps during loading and 2 steps during unloading. Figure 3 shows the hoop stress distribution at different t steps (t = 1, 6, 11, 13) where t is a time-like parameter for the purpose of bookkeeping. Figure 4 shows the distributions of residual radial and axial stresses and equivalent plastic strain. The residual stresses are considered to be elastic according to these two models. The unloading process may not be purely elastic if other models [4] are used. Future work should search for a more realistic model including the Bauschinger effect in a high strength steel [15]. Experimental measurements, if available, should be used for comparison with numerical predictions.

The tube of wall ratio 3(a = 1", b = 3") has also been loaded to reach fully plastic state and then unloaded completely. We use 11 steps during loading and 4 steps during unloading. Figure 5 shows the boundary displacements (U_a and U_b) as functions of pressure history. The numerical results for the displacements during unloading are very close between the two models. However, there are noticeable differences in the size of reverse yielding and the stresses within a small zone near the bore. There are 60 stations along the radial direction at which the stresses are calculated. At the end of complete unloading, reverse yielding occurs at 3 or 7 stations near the bore according to isotropic or kinematic models, respectively. Figure 6 shows the stresses at a point near the bore as functions of pressure history. The differences between the two models for the hoop and axial stresses during unloading are not small as can be seen in the figure.

Finally, the autofrettage solution for a closed volume chemical "bomb", is obtained for a tube with a = 0.865" and b = 4.005". The tube is loaded to p = 250 Ksi in 10 steps and then unloaded completely in 5 steps. At maximum pressure, 26 of its 60 stations have become plastic. At the end of complete unloading, reverse yielding occurs at 2 or 5 stations near the bore according to isotropic or kinematic models, respectively. Figure 7 shows the boundary displacements (U_a and U_b) as functions of pressure history. There are small differences for the displacements during unloading based on two models. The results for the stresses within the inner half of the tube are presented in Figures 8 and 9. Figure 8 shows the hoop stress at different stages of loading and unloading. Three stages (t = 1, 10, 15) represent the stage corresponding to initial yielding, maximum loading, and complete unloading, respectively. The differences for the hoop stresses during unloading based on two hardening models, are not small as can be seen in this figure. Figure 9 shows the differences for the axial and radial stresses within the inner half of the tube after complete unloading.

REFERENCES

1. R. Hill, Mathematical Theory of Plasticity, Oxford University Press, 1950.

2. P. G. Hodge and G. N. White, "A Quantitative Comparison of Flow and Deformation Theory of Plasticity," J. Appl. Mech., Vol. 17, pp. 180-184, 1950.

3. M. C. Steele, "Partially Plastic Thick-Walled Cylinder Theory," J. Appl. Mech., Vol. 19, pp. 133-140, 1952.

4. A. P. Parker, K. A. Sleeper, and C. P. Andrasic, "Safe Life Design of Gun Tubes - Some Numerical Methods and Results," Proc. of the 1981 Army Numerical Analysis and Computer Conference, pp. 311-333.

5. P. C. T. Chen, "The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes," Proc. Army Symposium on Solid Mechanics, The Role of Mechanics in Design - Ballistic Problems, pp. 243-253, 1972.

6. S. C. Chu, "A More Rational Approach to the Problem of an Elastoplastic Thick-Walled Cylinder," J. Franklin Institute, Vol. 294, pp. 57-65, 1972.

7. P. C. T. Chen, "Generalized Plane-Strain Problems in an Elastic-Plastic Thick-Walled Cylinder," Trans. 26th Conference of Army Mathematicians, pp. 265-275, 1980.

8. J. T. Oden, Finite Elements of Nonlinear Continua, McGraw-Hill, 1972.

9. O. C. Zienkiewicz, The Finite Element Method, 3rd Edition, McGraw-Hill, 1977.

10. A. K. Noor, "Survey of Computer Programs for Solution of Nonlinear Structural and Solid Mechanics Problems," Computer and Structures, Vol. 13, pp. 425-465, 1981.

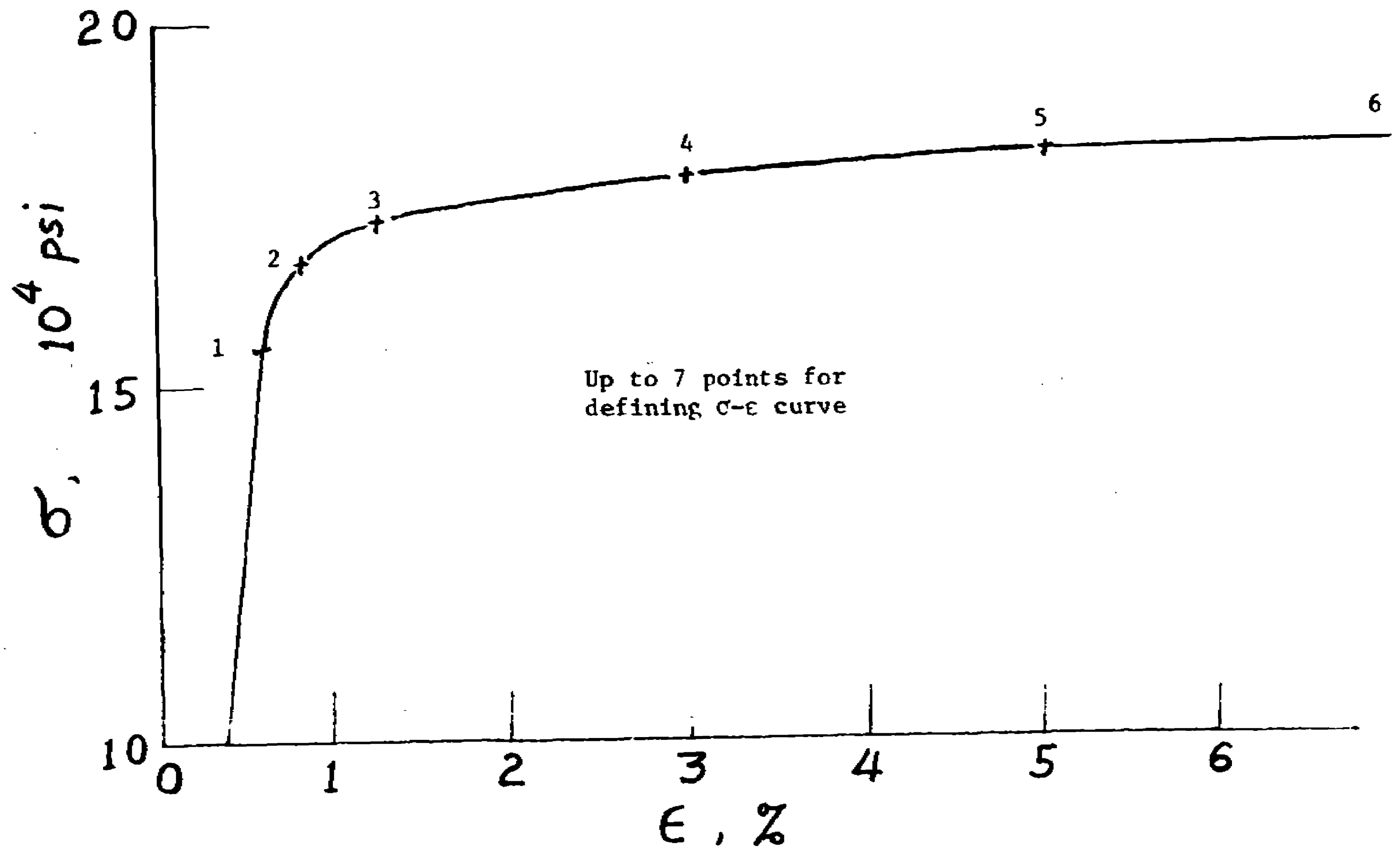
11. K. J. Bathe, "ADINA Users' Manual," Report AE-81-1, ADINA Engineering Inc., Watertown, MA, 1981.

12. K. J. Bathe, "ADINA Handbook - Theoretical Basis and Modeling Guide," Report AE-82-4, ADINA Engineering Inc., Watertown, MA. To appear in 1983.

13. W. Prager, "The Theory of Plasticity: A Survey of Recent Achievements," Proceedings of the Institution of Mechanical Engineers, Vol. 169, pp. 41-57, 1955.

14. H. Ziegler, "A Modification of Prager's Hardening Rule," Q. Appl. Math., Vol. 17, pp. 55-65, 1959.

15. R. V. Milligan, W. H. Koo, and T. E. Davidson, "The Bauschinger Effect in a High Strength Steel," J. Basic Engineering, Vol. 88, pp. 480-488, 1966.



Up to 7 points for defining σ-ε curve

Figure 1. Stress-Strain Curve for a High Strength Steel.

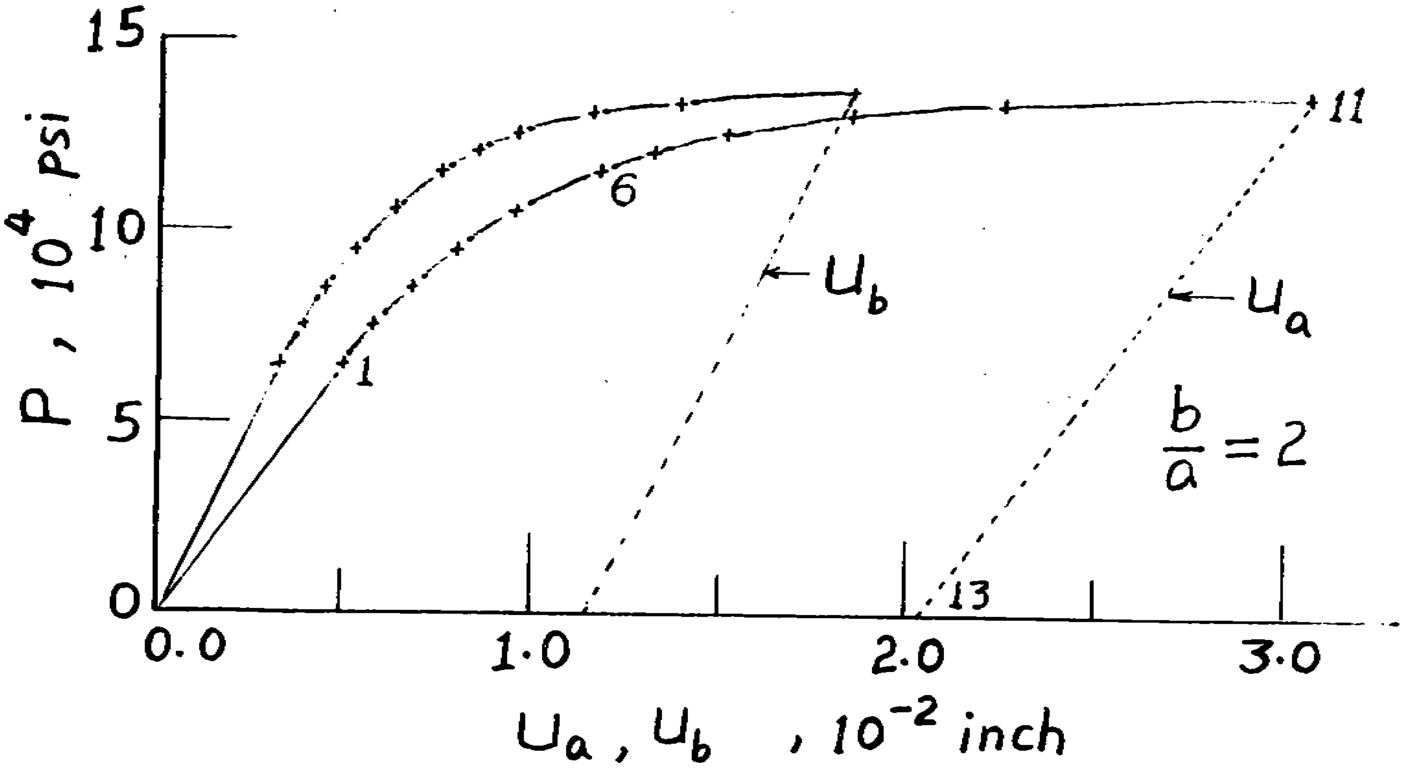


Figure 2. Boundary Displacements as Functions of Pressure History.

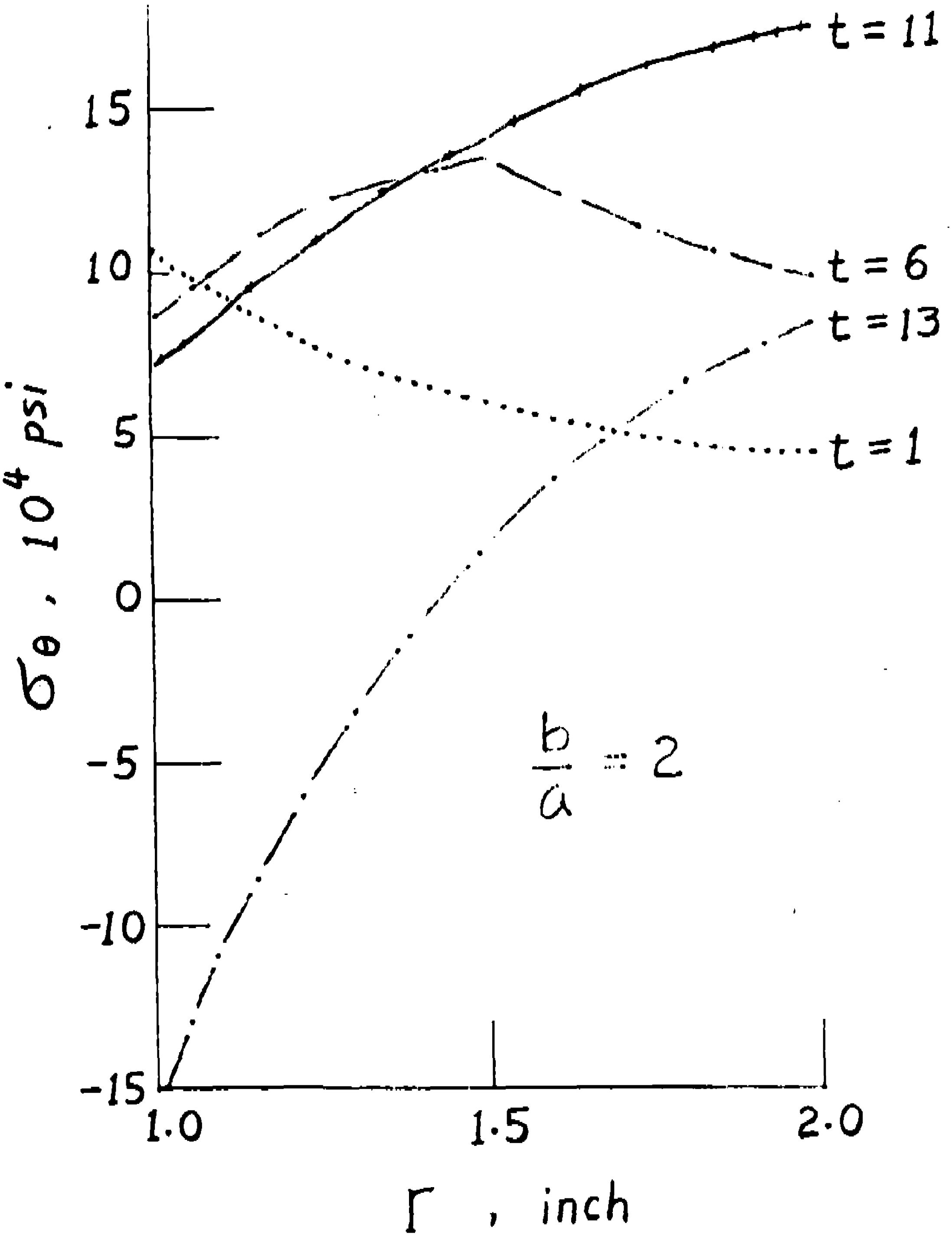


Figure 3. Hoop Stress Distribution at Different Stages of Loading.

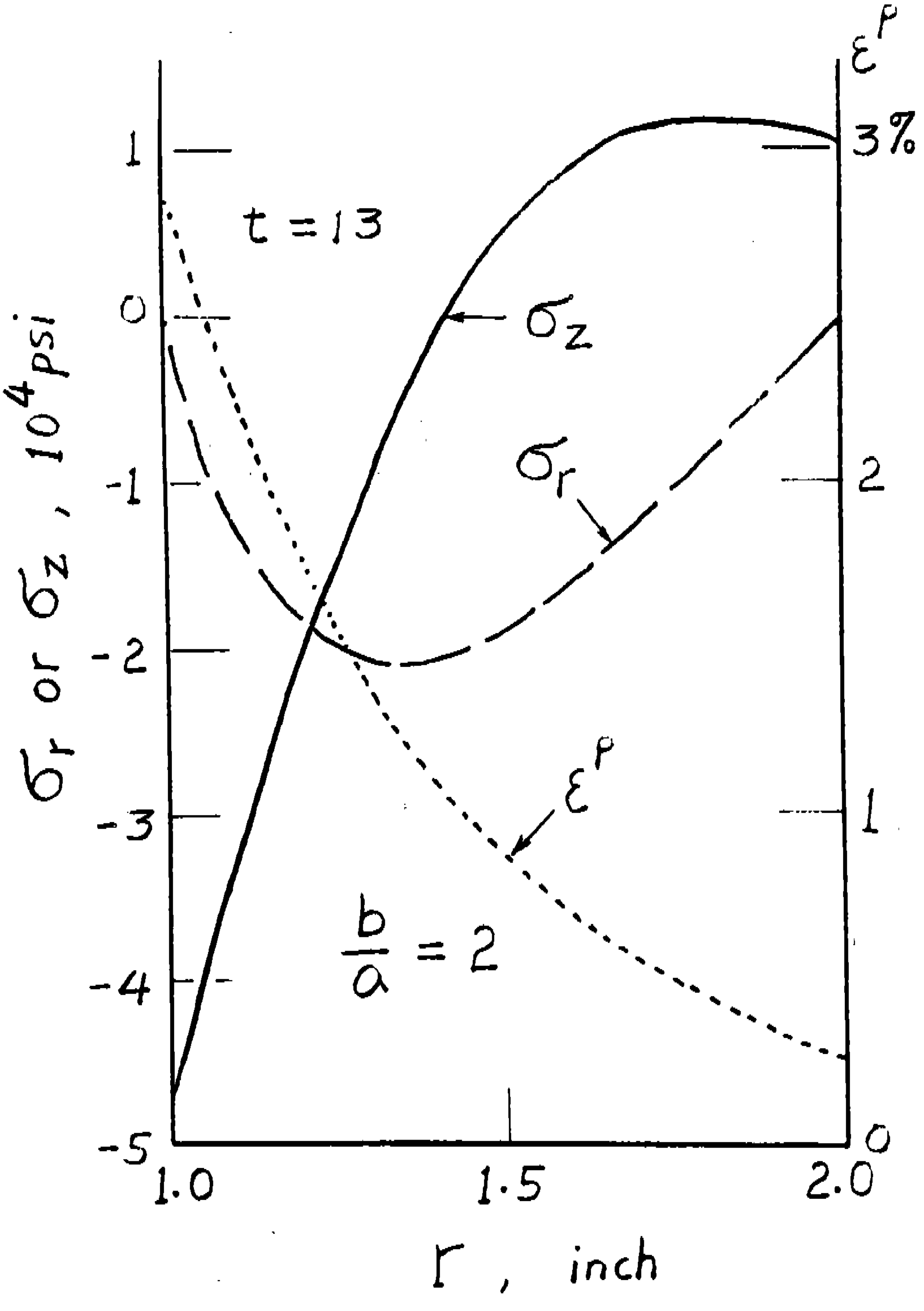


Figure 4. Distribution of Radial and Axial Stresses and Equivalent Plastic Strain After Complete Unloading.

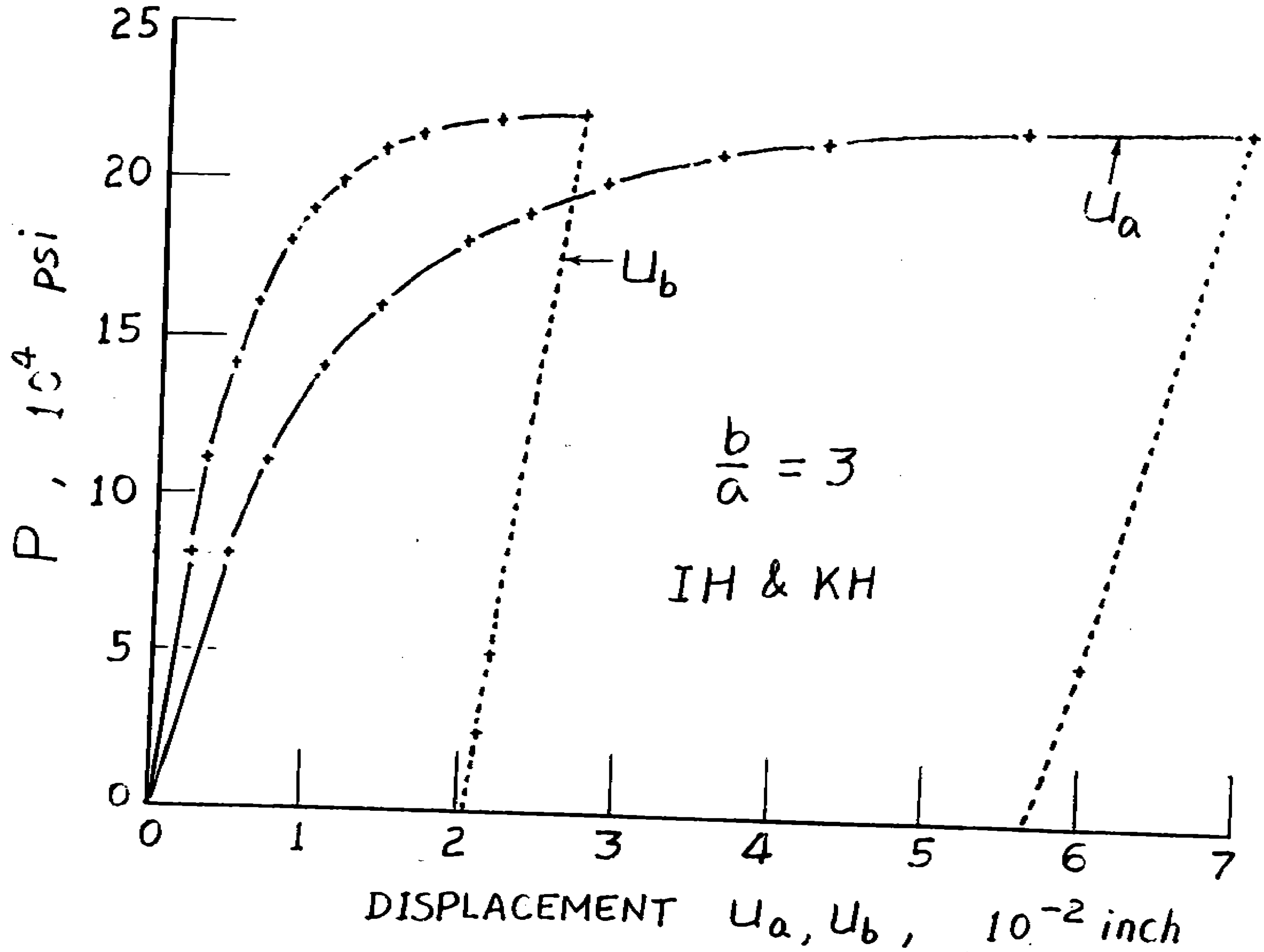


Figure 5. Boundary Displacements as Functions of Pressure History.

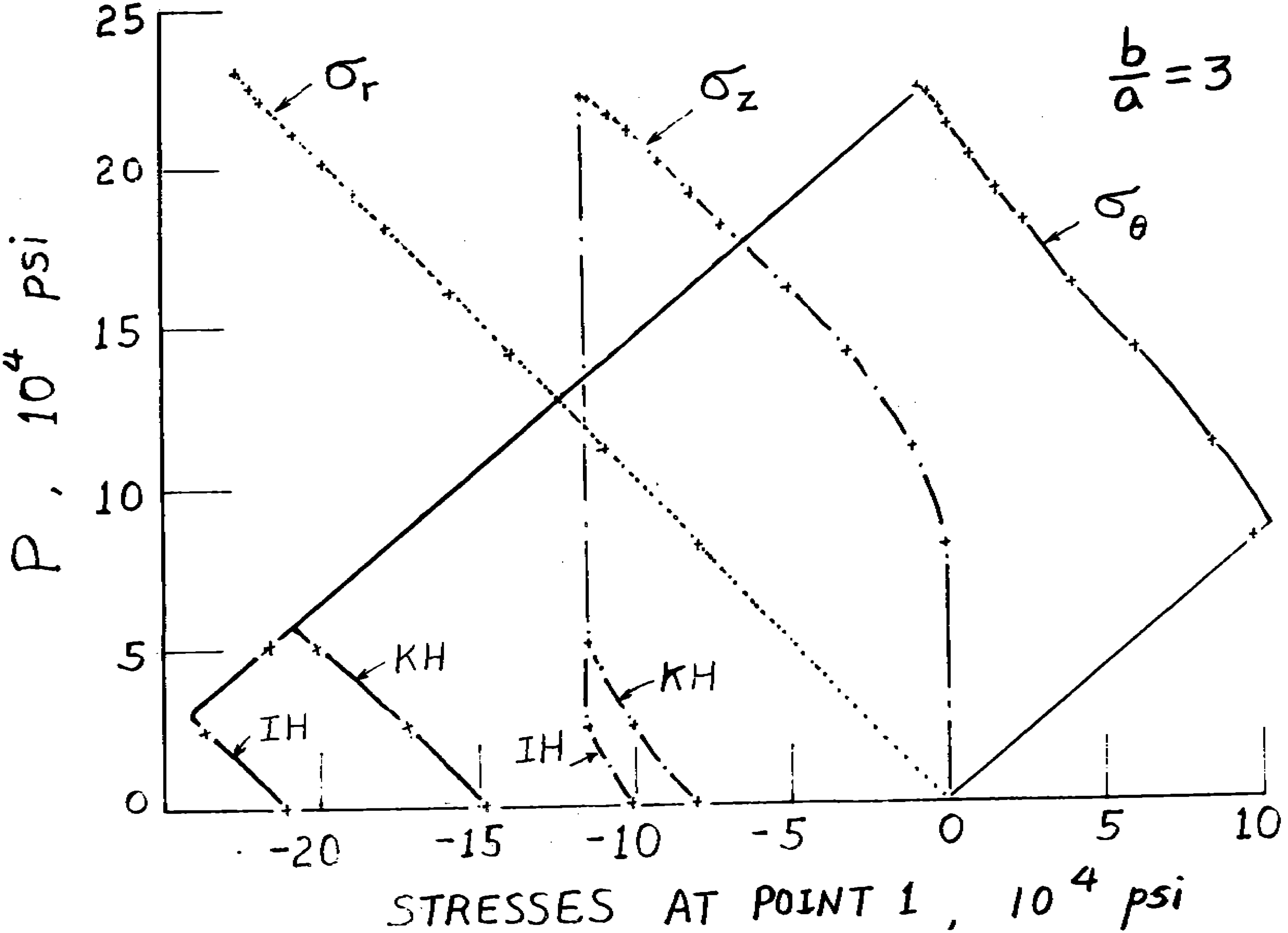


Figure 6. Stresses Near the Bore as Functions of Pressure History.

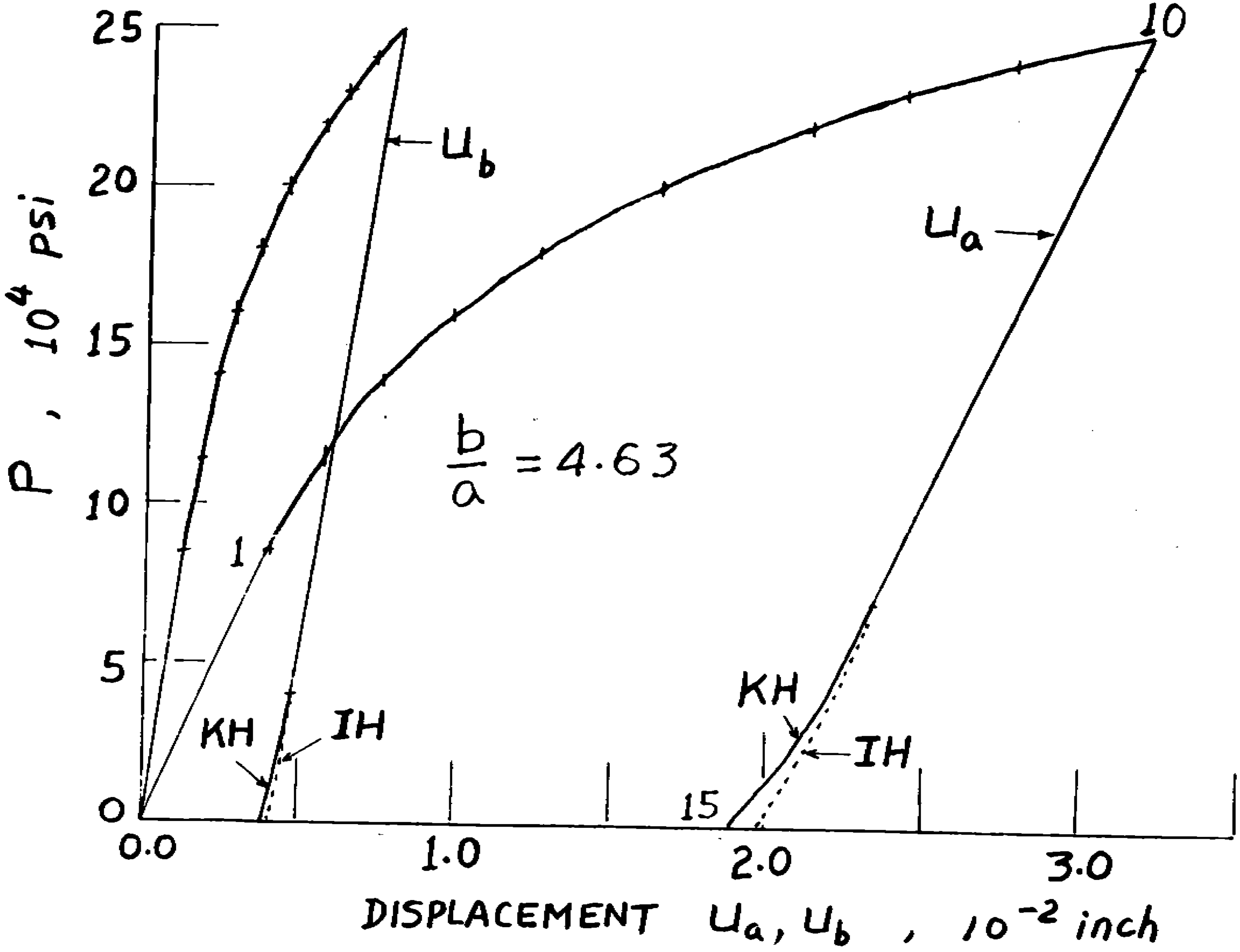


Figure 7. Boundary Displacements as Functions of Pressure History.

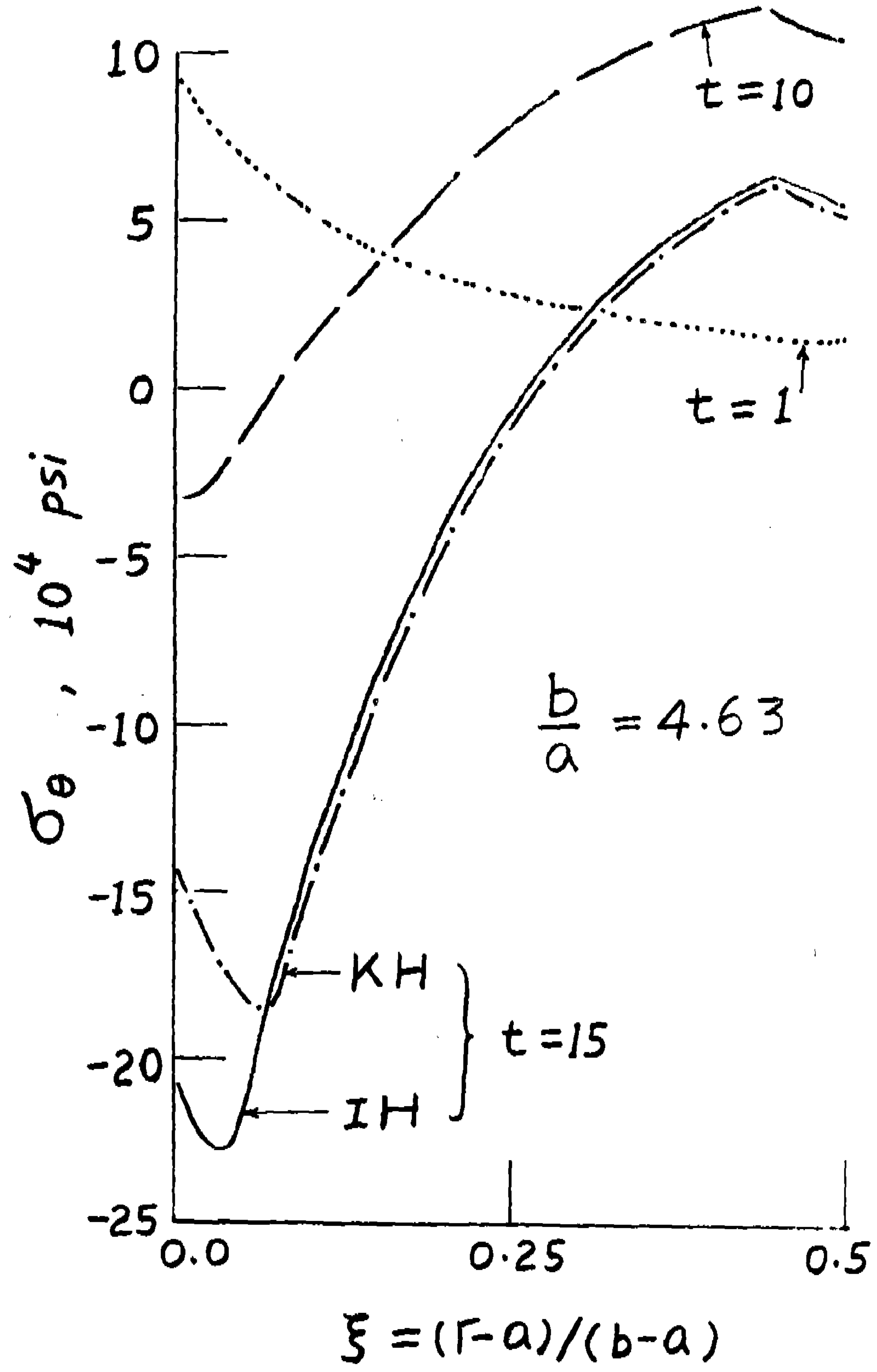


Figure 8. Distribution of Hoop Stress in the Inner Half at Different Stages of Loading.

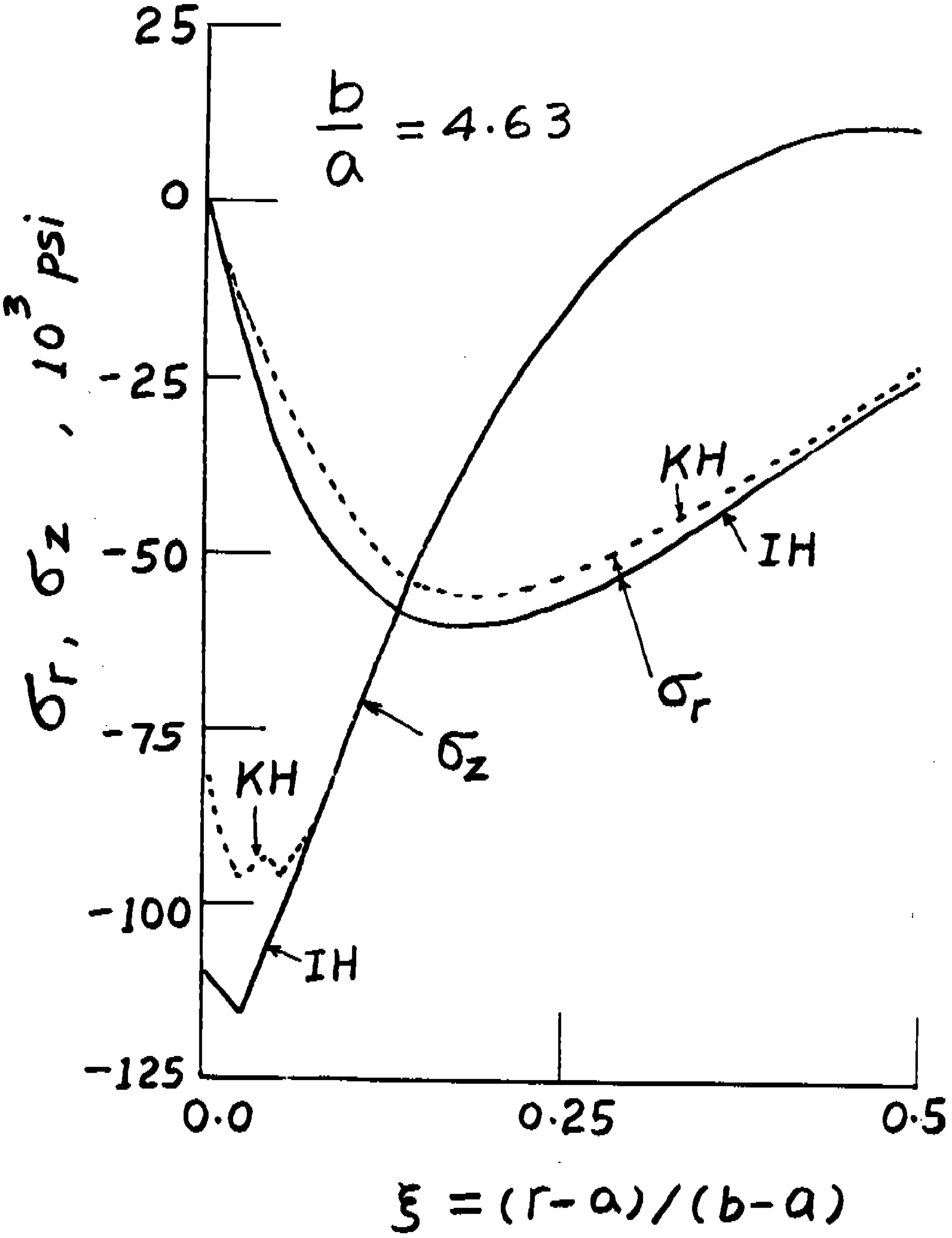


Figure 9. Distributions of Residual Radial and Axial Stresses in the Inner Half.

A FINITE INTEGRAL TRANSFORM (FIT) METHOD FOR DIFFERENTIAL EQUATIONS HAVING ASYMPTOTIC SOLUTIONS

Charles J. Daly
Earth Sciences Branch
Research Division
Cold Regions Research and Engineering Laboratory
Hanover, New Hampshire 03755

ABSTRACT. An analytical method is developed for differential equations whose solutions y(t) decay asymptotically for large t. The approach, called the finite integral transform (FIT) method, is based upon the use of selected transform-inverse pairs for which the inverses are known explicitly. Using the method, approximate analytical expressions for y(t) are obtained in the form of series of orthogonal functions $\psi_n(t)$. Asymptotic behavior of the approximating series is guaranteed by requiring each ψ_n to be asymptotic for large t. The central feature of the FIT approach is the generation of a system of algebraic equations which is solved for the vector of coefficients appearing in the orthogonal series solution. The FIT method amounts to a discretization of the unknown solution in a frequency domain. This is in contrast to finite difference and finite element methods which discretize over the domain of independent variables. The FIT approach is demonstrated by application to a simple linear ODE with variable coefficient.

BACKGROUND. The theory of finite integral transforms is well known, having its genesis in Fourier's analyses of the differential equations of heat flux. As Sneddon (1972) points out, finite transforms arise from the solutions of well-posed differential equations of the Sturm-Liouville variety. Since the different eigenfunctions corresponding to a particular Sturm-Liouville problem are mutually orthogonal, simple, yet extremely useful finite integral transform-inverse pairs can be formulated. For example, the problem:

$$\frac{d^2\phi}{dx^2} + \lambda^2\phi = 0 \qquad \phi(0) = \phi(\pi) = 0 \tag{1}$$

gives rise to the familiar finite sine transform-inverse pair:

$$G(n) = \mathcal{F}\left(g(x);\ x\to n\right) = \int_0^\pi g(x)\ \sin nx\ dx \tag{2}$$

$$g(x) = \frac{2}{\pi}\sum_{n=1}^{\infty} G(n)\ \sin nx\quad . \tag{3}$$

Other examples are the finite cosine, Hankel, Legendre, Tchebycheff, Mellin, and Laguerre transforms cited by Sneddon (1972).

Finite integral transforms, in general, have a remarkable and extremely useful property. Consider f(x) to be a piecewise continuous function for which the integral of a finite transform exists for all n. The enumerable set of values F(n) (the transform amplitudes) amounts to a complete description of f(x) over the interval of definition of the transform, excepting discontinuities. This observation is the basis of the FIT method for solving differential equations that are not amenable to the techniques of classical analysis. For example, given that the solution to a differential equation g(x) satisfies the boundary condition of (1), then there must be a finite sine representation of g of the form of (3); the problem becomes one of determining the values G(n). Using this approach, Daly and Morel-Seytoux (1981) were able to solve linear elliptic, and time invariant parabolic PDE's in terms of finite sine, cosine, and modified Laguerre transform series.

ANALYSIS. The Laguerre polynomials:

$$\mathcal{L}_n(t) = \frac{e^t}{n!}\frac{d^n}{dt^n}\left(t^n e^{-t}\right) \tag{4}$$

are the basis of a transform-inverse-pair described by McCully (1960) and Hirschman (1963):

$$A(n) = \mathcal{F}\left(a(t);\ t\to n\right) = \int_0^\infty e^{-t}\,\mathcal{L}_n(t)\ a(t)\ dt \tag{5}$$

$$a(t) = \sum_{n=0}^{\infty} A(n)\ \mathcal{L}_n(t) \qquad t > 0 \tag{6}$$

From a practical standpoint, the usefulness of a transform-inverse pair depends on the ability to approximate the inverse, for all t, by a finite number of terms. Considering (6) it is easy to demonstrate that any partial sum of two or more terms diverges as t becomes large. A modified Laguerre transform which is more appropriate for representing asymptotic functions is:

$$Y(n) = \mathcal{F}\left(y(t);\ t\to n\right) = \int_0^\infty e^{-t/2}\,\mathcal{L}_n(t)\ y(t)\ dt \tag{7}$$

$$y(t) = e^{-t/2}\sum_{n=0}^{\infty} Y(n)\ \mathcal{L}_n(t) \qquad t > 0 \tag{8}$$

To illustrate the FIT method consider the simple linear ODE:

$$\frac{dy}{dt} + a(t)\ y = b(t) \qquad y(0) = \xi \tag{9}$$

such that y decays asympotically as t becomes large. Assume that a(t) is given by M+1 terms of (6) so that (9) becomes:

$$\frac{dy}{dt} + y \sum_{m=0}^{M} A(m) \mathcal{L}_m(t) = b(t) \qquad y(0) = \xi \tag{10}$$

Taking the modified Laguerre transform of (10) according to (7) gives:

$$\begin{aligned} & -2\xi + Y(n) + 2 \sum_{p=0}^{n-1} Y(p) \\ & \quad + 2 \sum_{m=0}^{M} A(m) \int_0^\infty e^{-t/2} \mathcal{L}_n(t) \mathcal{L}_m(t)\, y(t)\, dt \\ & \quad = 2 \int_0^\infty e^{-t/2} \mathcal{L}_n(t)\, b(t)\, dt \equiv 2\, B(n) \end{aligned} \tag{11}$$

where the transform of the derivative is obtained by partial integration (Daly, 1979).

The product of two Laguerre polynomials can be written as a sum of Laguerre polynomials:

$$\mathcal{L}_m(t) \mathcal{L}_n(t) = \sum_{k=0}^{n+m} R(k;\, n,m) \mathcal{L}_k(t) \tag{12}$$

The coefficients R are derived in the Appendix. Substitution of (12) in (11) gives:

$$Y(n) + 2 \sum_{p=0}^{n-1} Y(p) + 2 \sum_{m=0}^{M} A(m) \sum_{k=0}^{n+m} R(k;n,m) Y(k) = 2\big(\xi + B(n)\big) \tag{13}$$

For n = 0,1,2,...N, (13) amounts to a set of linear algebraic equations for the N+1 values of Y(n). Assuming that y(t) is adequately represented by those N+1 coefficients:

$$y(t) \approx e^{-t/2} \sum_{n=0}^{N} Y(n) \mathcal{L}_n(t) = y_c(t) \tag{14}$$

A Gauss-Seidel iterative procedure can be used to solve (13) for the Y(n).

As an example, consider the simple problem:

$$\frac{dy}{dt} + 2ty = (4t-1)e^{-t/2} \quad y(0) = 10 \tag{15}$$

which can be solved for the exact solution:

$$y_e(t) = 8e^{-t^2} + 2e^{-t/2} \tag{16}$$

Using the FIT approach, where it is assumed that y(t) can be adequately represented by N+1 = 21 terms, (15) is solved for the spectral amplitudes Y(n) shown in Figure 1. Amplitudes are calculated to five digit accuracy by the Gauss-Seidel procedure.

Figure 2 is a comparison of $y_e(t)$ and $y_c(t)$ given by (16) and (14) respectively. The FIT solution is an excellent fit to the closed form solution.

CONCLUSIONS. The finite integral transform method has been demonstrated for the solution of a linear ODE with variable coefficient. The apparent success of the procedure is an indication of the potential of FIT methods for solving more complex problems. Advantages of the approach include the generation of analytic expressions and the lack of node or grid systems as required by many alternative numerical methods.

APPENDIX. Because the Laguerre polynomials are orthogonal on $(0,\infty)$ with respect to e^{-t}, (12) becomes

$$R(k;n,m) = \int_0^\infty e^{-t} \mathcal{L}_k(t)\, \mathcal{L}_n(t)\, \mathcal{L}_m(t)\, dt \tag{A1}$$

Expanding the product of Laguerre polynomials gives:

$$\mathcal{L}_n(t)\, \mathcal{L}_m(t) = \sum_{p=0}^{n} \frac{(-1)^p}{p!} \binom{n}{p} t^p \sum_{r=0}^{m} \frac{(-1)^r}{r!} \binom{m}{r} t^r$$

$$= \sum_{p=0}^{n+m} \sum_{\substack{i=0\\ i<n\\ i\geq p-m}}^{p} \frac{(-1)^p}{i!(p-i)!} \binom{}{i} \binom{m}{p-i} t^p \tag{A2}$$

Let:

$$c(p) = \sum_{\substack{i=0\\ i<n\\ i\geq p-m}}^{p} \frac{(-1)^p}{i!(p-i)!} \binom{n}{i} \binom{m}{p-i} \tag{A3}$$

Then:

$$R(k;n,m) = \sum_{p=0}^{n+m} c(p) \int_0^\infty e^{-t} \mathcal{L}_k(t)\, t^p dt \tag{A4}$$

Evaluating the integral of (A4) and substituting (A3) gives:

$$R(k;n,m) = \sum_{\substack{p=0\\ p\geq k}}^{n+m} c(p)\, (-1)^k \binom{p}{k} p!$$

$$= \sum_{p=k}^{n+m} (-1)^{p+k} \binom{p}{k} \sum_{\substack{i=0\\ i<n\\ i\geq p-m}}^{p} \binom{n}{i} \binom{m}{p-i} \binom{p}{i} \tag{A5}$$

REFERENCES

Daly, C.J. (1979) Analytical/numerical methods for groundwater flow and quality problems. Ph.D. dissertation, Colorado State University, Fort Collins.

Daly, C.J. and H.J. Morel-Seytoux (1981) An integral transform method for the linearized Boussinesq groundwater flow equation. Water Resources Research, v. 17, no. 4: 875-884.

Hirshman, I.I. (1963) Laguerre transforms. Duke Mathematics Journal, v. 30: 495-10.

McCully, J. (1960) The Laguerre transform. SIAM Review, v. 2: 185-191.

Sneddon, I.N. (1972) The Use of Integral Transforms. McGraw-Hill: 539 pp.

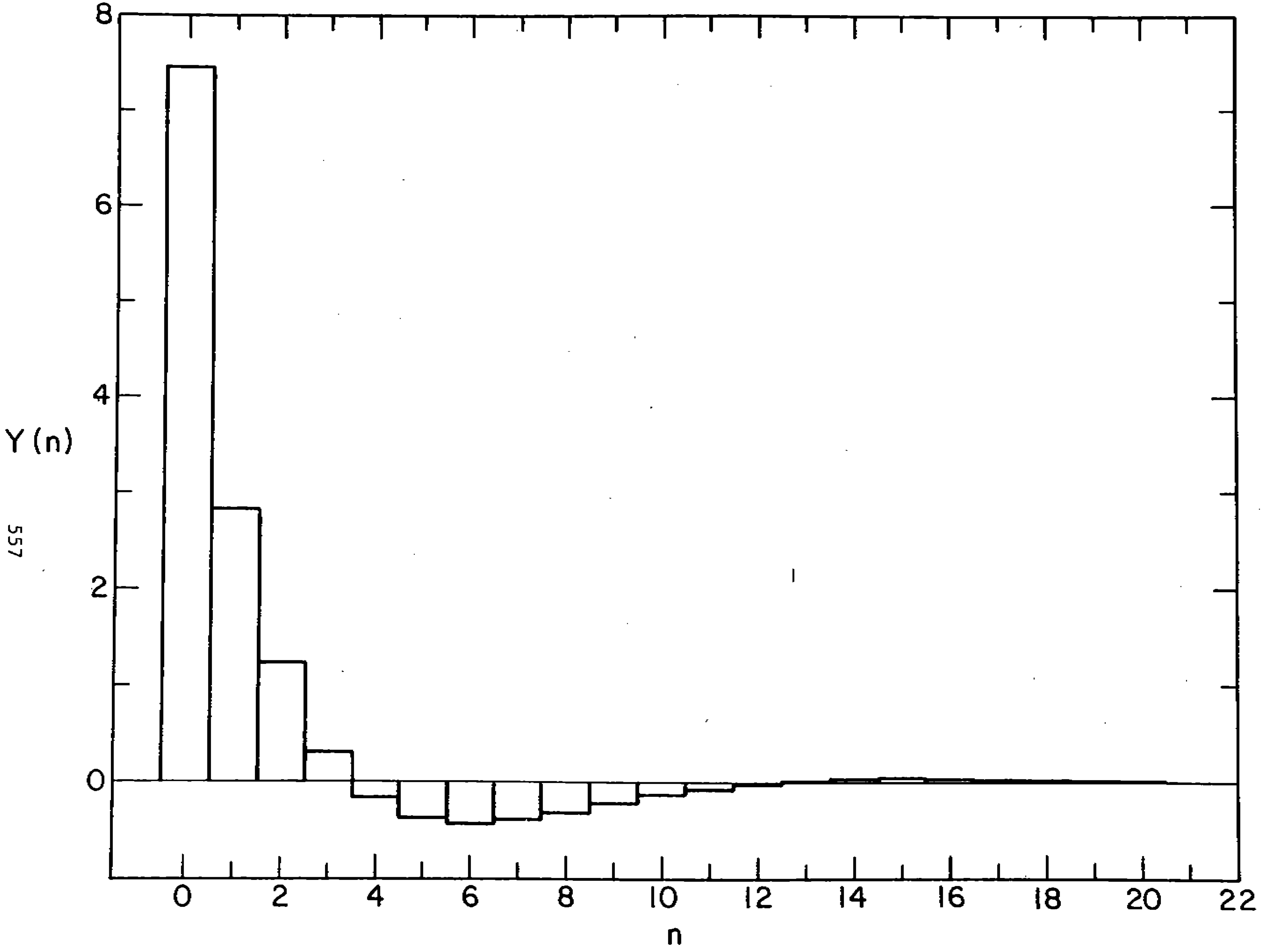


FIGURE 1

Spectral amplitudes as determined from solution of (13)

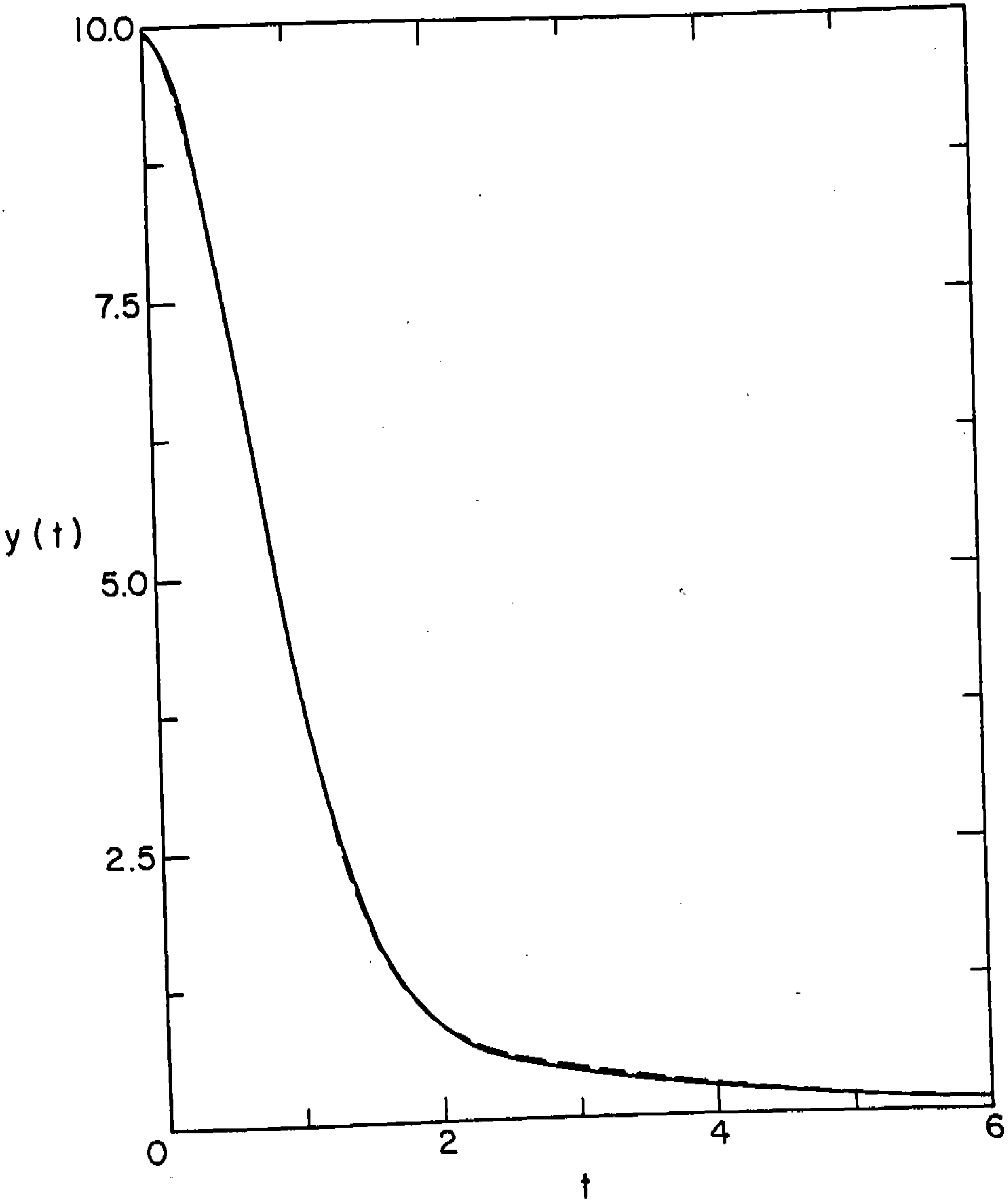


FIGURE 2

$y_e(t)$ (dashed line); $y_c(t)$ (solid line)

A FINITE DIFFERENCE METHOD FOR ANY PARTIAL DIFFERENTIAL EQUATION

R. YALAMANCHILI

TECHNOLOGY BRANCH, ARMAMENT DIVISION
FIRE CONTROL & SMALL CALIBER WEAPON SYSTEMS LABORATORY
U.S. ARMY ARMAMENT RESEARCH AND DEVELOPMENT CENTER, DOVER, NJ 07801

ABSTRACT

The most general heat diffusion equation possesses not only first- and second-order derivatives in space but also first-and second-order derivatives in time. Therefore, the governing equation can be parabolic, elliptic or even hyperbolic depending upon the parameters chosen. The model includes various physical problems, such as, steady and unsteady classical heat conduction (usually known as classical Fourier model), heat pulse (Non-Fourier model), abrasive cut-off and surface grinding operations in machining of metal components. An explicit and unconditionally stable finite difference scheme is developed for the general purpose governing equation. The heat transfer example is included to discuss the accuracy and stability of this numerical scheme.

I. INTRODUCTION.

The mathematical model, considered here, represents various physical problems in the area of heat transfer. These include numerous problems in steady and unsteady heat conduction; large heat flux applications, such as, plasma torch and hypervelocity weapon systems; and problems arising from manufacturing operations, such as, abrasive cut-off and surface grinding. Although all these physical problems look entirely different, one may be able to generalize them into a single model and prepare a single numerical method for its solution. It is important to realize that the generalized model can be specialized to parabolic, elliptic or even hyperbolic depending upon the problem of interest.

Currently the finite-difference methods are common for the solution of partial differential equations in both design and research and development because of the advent of super and inexpensive computers. The numerical methods for the unsteady heat diffusion equation receive most attention not only because of heat transfer applications but also because they are the basis for any study of parabolic partial differential equations. It is common to test the heat conduction example first even though the methods are developed for complex nonlinear parabolic partial differential equations which may arise in other fields. Yalamanchili [1, 2] showed that the finite-element and finite-difference methods belong to the class of method of weighted residuals. It is also concluded that the finite-element method is more conservative in both stability and nonoscillation characteristics than the finite-difference method, but not as conservative as the method of weighted residuals. Since the finite-element method is unique because of Gurtin's [3] variational principle and numerous finite-differences can be constructed with ease, it is found that some finite-difference schemes are better than the finite-element scheme in accuracy also. Therefore, further attention is focussed here on finite-difference schemes only. An example is considered to show where the present solution method stands in comparison to similar numerical solution methods.

II. GENERALIZED HEAT CONDUCTION MODEL.

The generalized model including charring ablation [4] may be stated as

$$\frac{\partial T}{\partial t} - u\frac{\partial T}{\partial t} - v\frac{\partial T}{\partial t} + \frac{1}{a^2}\frac{\partial^2 T}{\partial t^2} = \frac{1}{r^j}\frac{\partial}{\partial r}\left(\alpha r^j \frac{\partial T}{\partial r}\right) + \frac{\partial}{\partial z}\left(\alpha \frac{\partial T}{\partial z}\right) \qquad (1)$$

where T = Temperature

t = time

u = speed of tool or work piece in the radial direction

v = speed of tool or work piece in the axial direction

r = radial coordinate

z = axial coordinate

j = 0, plane

= 1, axisymmetric body

α = thermal diffusivity

a = speed of sound

The following dimensionless variables are introduced for dependent and independent variables:

$$\beta = aR/\alpha \ , \ \gamma = az/\alpha \ , \ \delta = a^2t/\alpha \ , \text{ and } \theta = (T - T_o)/(T_w - T_o) \tag{2}$$

where R = r for a plane

= ℓnr for axisymmetric case

T_o = Initial temperature

T_w = wall temperature

The resulting governing equation in terms of dimensionless variables is shown as

$$P \frac{\partial^2\theta}{\partial\delta^2} + Q \frac{\partial\theta}{\partial\delta} = D \frac{\partial^2\theta}{\partial\beta^2} + C \frac{\partial\theta}{\partial\beta} + H \frac{\partial^2\theta}{\partial\gamma^2} + G \frac{\partial\theta}{\partial\delta} \ . \tag{3}$$

P equal to unity represents a heat pulse (large heat flux) problem. Otherwise, it is zero. Q equals unity for any transient case. C is one for abrasive cut-off, whereas G is unity for surface grinding. Otherwise, C and G will be zero. The value of D will depend upon whether plane or axisymmetric, as defined before. H will be zero for one-dimensional problems. Otherwise, it is one.

III. NUMERICAL METHOD.

The dimensionless governing equation is not only general from the physical point of view, but also mathematically. This can be reduced to parabolic, elliptic, or hyperbolic by appropriate selection of parameters P, Q, D, C, H, and G. It is quite common to use an entirely different method depending upon its mathematical characteristics. However, a simple explicit finite-difference formulation is utilized to obtain the solution numerically.

The following difference notation is utilized:

$\Theta = \Theta(\beta,\gamma,\delta) = \Theta_{i,j,k}$

$\Delta\delta_n$ = new time-step

$\Delta\delta_o$ = previous time-step

$$\text{Let } f(\delta) = a + b\delta + c\delta^2 \ . \tag{4}$$

Choose δ equal to zero at a time when the old time step $\Delta\delta_o$ has been changed to new, $\Delta\delta_n$. Let the corresponding time subscripts be k-1, and k+1. The coefficients of the quadratic equation become

$$\begin{aligned} a &= f_k \\ b &= \frac{f_{k+1} - f_k}{\Delta\delta_n} - \frac{f_{k-1}\Delta\delta_n + f_{k+1}\Delta\delta_o - (\Delta\delta_n + \Delta\delta_o)f_k}{\Delta\delta_n(\Delta\delta_o + \Delta\delta_o)} \\ c &= \frac{f_{k-1}\Delta\delta_n + f_{k+1}\Delta\delta_o - (\Delta\delta_o + \Delta\delta_n)f_k}{\Delta\delta_n\Delta\delta_o(\Delta\delta_n + \Delta\delta_o)} \quad . \end{aligned} \tag{5}$$

The derivatives at time, k, can be written as

$$\begin{aligned} \frac{\partial f}{\partial \delta} &= b \\ \frac{\partial^2 f}{\partial \delta^2} &= 2c \\ \frac{\partial f}{\partial \delta} &= \frac{f_{k+1} - f_k}{\Delta\delta_n} \qquad \text{if forward differences are utilized.} \end{aligned} \tag{6}$$

In addition to these time derivatives, if central differences are utilized for second order spatial derivatives, the governing equation in difference format may be written as:

$$\begin{aligned} &\left[\frac{2P}{\Delta\delta_n(\Delta\delta_n + \Delta\delta_o)} + \frac{Q}{\Delta\delta_n}\right]\Theta_{i,j,k+1} = \left[\frac{2P}{\Delta\delta_n\Delta\delta_o} + \frac{Q}{\Delta\delta_n} - \left(\frac{2D}{\Delta\beta^2} + \frac{C}{\Delta\beta} + \frac{2H}{\Delta\delta^2} + \frac{G}{\Delta\gamma}\right)\right]\cdot \\ &\cdot\Theta_{i,j,k} - \frac{2P}{\Delta\delta_o(\Delta\delta_n + \Delta\delta_o)}\Theta_{i,j,k-1} + \left(\frac{D}{\Delta\beta^2} + \frac{C}{\Delta\beta}\right)\Theta_{i+1,j,k} \\ &+ \frac{D}{\Delta\beta^2}\Theta_{i-1,j,k} + \left(\frac{H}{\Delta\gamma^2} + \frac{G}{\Delta\gamma}\right)\Theta_{i,j+1,k} + \frac{H}{\Delta\gamma^2}\Theta_{i,j-1,k} \quad . \end{aligned} \tag{7}$$

IV. STABILITY ANALYSIS.

The difference equation, mentioned above, is not unconditionally stable. According to Dusinberre [5], based on laws of thermodynamics, this scheme is stable only if the following condition is satisfied:

$$\frac{2P}{\Delta\delta_n\Delta\delta_o} + \frac{Q}{\Delta\delta_n} \geq \frac{2D}{\Delta\beta^2} + \frac{C}{\Delta\beta} + \frac{2H}{\Delta\gamma^2} + \frac{G}{\Delta\gamma} \quad . \tag{8}$$

However, substitution of the following relationship into some of the terms of the above difference equation yiels an unconditionally stable difference equation.

$$\Theta_{i,j,k} = \frac{\Delta\delta_n \Theta_{i,j,k-1} + \Delta\delta_o \Theta_{i,j,k+1}}{\Delta\delta_o + \Delta\delta_n} . \qquad (9)$$

This relatio implies a linear variation of temperature between times subscripts (k-1) and (k+1). the resulting unconditionally stsble fifference equation is given below:

$$A\Theta_{i,j,k+1} = \left(\frac{2P}{\Delta\delta_n \Delta\delta_o} + \frac{Q}{\Delta\delta_n}\right)\Theta_{i,j,k} - B\Theta_{i,j,k-1} + \left(\frac{D}{\Delta\beta^2} + \frac{C}{\Delta\beta}\right)\Theta_{i+1,j,k}$$

$$+ \frac{D}{\Delta\beta^2}\Theta_{i-1,j,k} + \left(\frac{H}{\Delta\delta^2} + \frac{G}{\Delta\delta}\right)\Theta_{i,j+1,k} + \frac{H}{\Delta\delta^2}\Theta_{i,j-1,k} . \qquad (10)$$

$$\text{Where } A = \frac{2P}{\Delta\delta_n(\Delta\delta_n + \Delta\delta_o)} + \frac{Q}{\Delta\delta_n} + \left(\frac{2D}{\Delta\beta^2} + \frac{C}{\Delta\beta} + \frac{2H}{\Delta\delta^2} + \frac{G}{\Delta\delta}\right)\frac{\Delta\delta_o}{\Delta\delta_n + \Delta\delta_o} \qquad (11)$$

$$\text{and } B = \frac{2P}{\Delta\delta_o(\Delta\delta_n + \Delta\delta_o)} + \left(\frac{2D}{\Delta\beta^2} + \frac{C}{\Delta\beta} + \frac{H}{\Delta\delta^2} + \frac{G}{\Delta\delta}\right)\frac{\Delta\delta_n}{\Delta\delta_n + \Delta\delta_o} . \qquad (12)$$

The truncation error of the difference equation did not increase due to the assumption of linear variation of temperature within any two consecutive time-steps. Surprisingly, the truncation error is reduced somewhat due to the cancellation of some of the terms. The subitution of an assumption also elimates one deficiency exhibited by the Dufort and Frankel method [6] for nonlinear problems in that large variations in 'α' can be tolerated without introducing significant errors.

The unconditional stability can be shown by assuming an error, E, as a linear combination of terms of the form

$$\Theta_{i,j,k} = \eta^k e^{\sqrt{-1}\, i\Omega\Delta\beta} e^{\sqrt{-1}\, j\phi\Delta\gamma} \qquad (13)$$

and substituting into the final difference equation, one can obtain the relation for damping ratio (η) as

$$\eta^2 - \frac{E}{A}\eta + \frac{B}{A} = 0 . \qquad (14)$$

Where $$E = \frac{2P}{\Delta\delta_n \Delta\delta_o} + \frac{Q}{\Delta\delta_n} + \left(\frac{D}{\Delta\beta^2} + \frac{C}{\Delta\beta}\right)e^{\sqrt{-1}\Omega\Delta\beta} + \frac{D}{\Delta\beta^2}\, e^{-\sqrt{-1}\Omega\Delta\beta}$$

$$+ \left(\frac{H}{\Delta\gamma^2} + \frac{G}{\Delta\gamma}\right)e^{\sqrt{-1}\phi\Delta\gamma} + \frac{H}{\Delta\gamma^2}\, e^{-\sqrt{-1}\phi\Delta\gamma} \quad . \qquad (15)$$

It can be shown that the magnitude of the damping ratio (for amplitude), η , will be less than or equal to unity for any chosen parameters. Therefore, the final difference equation will remain stable for any constant integration interval or when the new interval is not larger than the old step plus the critical step increment $\alpha\Delta\delta_o/(\Delta\beta^2 + \Delta\gamma^2)$.

V. PARABOLIC EXAMPLE.

An exact analytical solution [7] exists for the one-dimensional problem with zero initial temperature everywhere and length, L. The boundary conditions are a step change to a temperature of unity on one end and insulated (zero heat flux) on the other end. Instead of approximating the heat flux boundary condition by a difference equation, one can also consider a body of length 2L and apply a step change in temperature to unity on both ends. In this case, there is no error involved in consideration of zero heat fluxes or insulated surfaces. The typical computer time is 18 seconds for 40 spatial-steps and 600 time-steps. The results obtained by the present method as well as by several other methods are shown in Tables 1 and 2.

TABLE 1. Comparison of Errors ($\times 10^5$)

Method	X/L=0.2	X/L=0.6	X/L=0.8
Analytical	.080929	.371439	.654747
Crank-Nicholson[8]	121	244	175
Saul'yev[9]	218	252	231
Liu[10]	351	240	134
Dufort-Frankel	90	367	296
Present	194	96	35

TABLE 2. Comparison of Errors(X10^5)

Method	X/L=0.2	X/L-0.6	X/L=0.8
Analytical	.647367	.782053	.885416
Crank-Nicholson[8]	4.1	2.6	1.4
Saul'yev[9]	120	72	33
Liu[10]	72	45	26
Dufort-Frankel	126	79	42
Present	27	16	9

VI. CONCLUSIONS.

The Fourier number is 0.1 for the results in Table 1. The present method is more accurate than not only other stable explicit schemes but also the famous implicit scheme at dimensionless positions of 0.6 and 0.8. Similar results are shown in Table 2 for a later time (Fourier number) of 0.5. The present method is much better than any other stable explicit scheme for all positions. Only, the Crank-Nicholson implicit scheme exceeded the accuracy of the present approach for larger dimensionless times (but not for small times). However, the Crank-Nicholson scheme requires more computer time because of implicit procedures than the present method. Therefore, the generalized model and the numerical method are a great asset for scientists and engineers.

VII. REFERENCES.

[1] Yalamanchili, R., and Chu, S. C., "Stability and Oscillation Characteristics of Finite-Element, Finite-Difference, and Weighted-Residual Methods for Transient Two-Dimensional Heat Conduction in Solids, "J. of Heat Transfer, Transactions of ASME, May 1973.

[2] Yalamanchili, R., "Accuracy, Stability, and Oscillation Characteristics of Transient Two-Dimensional Heat Conduction, "ASME Paper #75-WA/HT-85, 1975 ASME Winter Annual Meeting, Houston, Texas, November 1975.

[3] Gurtin, M.E., "Variational Principles for Linear Initial Value Problems, "Quarterly J. of Applied Mathematics, Vol. 22, 1964, pp. 252-256.

[4] Friedman, H. A., and McFarland, B. L. , "Two-Dimensional Transient Ablation and Heat Conduction Analysis for Multi-Material Thrust Chamber Walls, "AIAA Journal, Vol. 5, #7, pp. 753-761 (1968).

[5] Dusinberre, G. M., "Heat Transfer Calculations by Finite Differences," Second Edition, International Text Book Co., Scranton, PA 1961.

[6] Dufort, E. C., and Frankel, S.P., "Stability Conditions in the Numerical Treatment of Parabolic Differential Equations," Math Tables and Aids to Computation, Vol. 7, 1953, pp. 135-152.

[7] Crank, J., "The Mathematics of Diffusion," Oxford Press, London, 1956, pp. 20-21.

VII. REFERENCES.

[8] Crank, J., and Nicholson, P., "A Practical Method for Numerical Evaluation of Partial Differential Equations of Heat Conduction Type," Proc. of the Cambridge Philosophical Society, Vol. 43, 1947, pp. 50-67.

[9] Saul'yev, V. K., "Integration of Equations of Parabolic Type by the Method of Nets," Macmillan Co., New York, 1964.

[10] Liu, S. L., "Stable Explicit Approximations to Parabolic Partial Differential Equations," AICHE Journal, Vol. 15, #3, 1969, pp. 334.

NUMERICAL BOUNDARY CONDITIONS FOR FLOW PROBLEMS

George J. Fix and Max D. Gunzburger
Department of Mathematics
Carnegie-Mellon University
Pittsburgh, PA 15213

ABSTRACT. The development, analysis, and implementation of numerical boundary conditions for flow calculations in infinite domains are discussed. Emphasis is placed on potential flow, periodic acoustic, and incompressible viscous flow problems. In all cases, the infinite domain problems are approximated by problems posed on a bounded domain. To close the numerical problem, an artificial numerical boundary condition is imposed. The effect of these approximate boundary conditions on the accuracy of the numerical computations is examined.

I. INTRODUCTION. When a physical problem is posed on an unbounded domain, its solution cannot be directly approximated by a numerical technique. There are four classes of methods employed to treat such problems. The first is to map, either numerically or analytically, the infinite domain into a finite one, and then discretizing the governing partial differential equations by standard techniques, e.g., finite difference or finite element methods. This class of methods has often proved to be quite successful, especially for potential flow problems.

However, for complicated regions, e.g., the exterior to an airfoil, the numerical generation of the mapping function may be rather costly. Furthermore, for problems with oscillatory solutions, e.g., periodic acoustic problems, mapping to a finite domain just trades the problem of the infinite domain for the equally impossible problem of resolving very rapidly (in fact, infinitely rapidly) oscillating solutions within the new bounded domain.

A second popular approach for exterior problems is to transform the partial differential equation problem on an infinite domain into an integral equation problem posed on the finite boundary of the domain. The integral equation may then be discretized by the usual techniques. This type of approach is limited to problems for which a free space Green's function is known. Furthermore, discretization of the integral equation leads to dense (but finite) matrix problems.

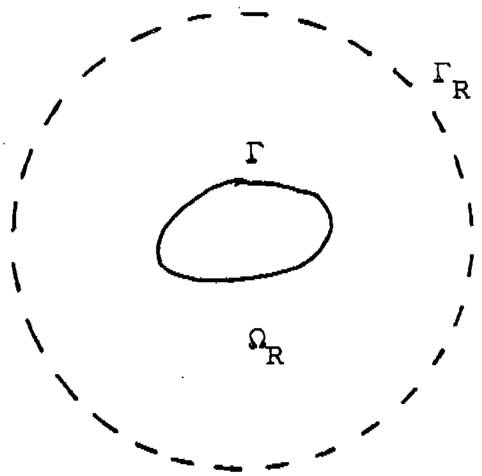
A third method of treating infinite domain problems is related to the integral equation approach. This approach couples a far field solution with a discretization by finite differences or finite elements in a bounded region. The coupling is effected by introducing auxiliary variables, defined on the artificial boundary of the truncated domain, and requiring these auxiliary variables to satisfy an integral equation along that boundary. This integral equation carries information about the solution in that part of the original domain lying outside the truncated domain. In order to apply this method, we need only know a free space Green's function for the differential operator governing the solution outside of the truncated domain. One way to view this method is to think of it as solving the partial differential equation problem on a bounded truncated domain, with the integral equation serving as an exact non-local boundary condition. By non-local we mean that the (unknown) solution at all the points on the artificial boundary are coupled. This method is effective but involves the introduction of additional unknowns and often results in complications in the resulting linear systems which must be solved.

The fourth method, which is the one considered in this paper, again solves the partial differential equation problem on a truncated domain. However, we now impose, on the artificial boundary, an approximate local boundary condition. By local, we mean that the boundary condition holds pointwise at the boundary. Since this approach involves only an approximate boundary condition, it is desirable to develop artificial boundary conditions of high accuracy. Indeed, the method can be cost effective only when the truncated domain is relatively small in its extent. However, when the method works, it is the simplest to implement, since it essentially consists of a direct discretization of the partial differential equations on a bounded domain with local boundary conditions, i.e., as far as the discretization technique, it is essentially one for an interior problem.

In Section II, we discuss potential flow, periodic acoustic, and incompressible viscous flow problems in exterior domains.

In Section III we discuss incompressible viscous flow problems in channels. We also refer the reader to [1] which was presented at a previous Army conference, which treats numerical boundary conditions for unsteady wave propagation problems.

II. FLOW PROBLEMS IN EXTERIOR DOMAINS. In this section we first consider problems governed by second order elliptic partial differential equations and which involve exterior domains, i.e., domains which are the exterior of a bounded body in R^2 or R^3. In the vicinity of the body, the "near field", the equation may have variable coefficients. However, away from the body, the "far field", we assume that the equation reduces to one with constant coefficients, at least in an asymptotic sense. We introduce an artificial boundary, denoted by Γ_R, and assume that it contains the inner boundary which is denoted by Γ. The original infinite domain we denote by Ω, and the truncated domain, i.e., the region between Γ and Γ_R, by Ω_R. We assume that the artificial boundary Γ_R is placed sufficiently far away from the inner boundary Γ so that the differential equation, in the vicinity of Γ_R, is already in its far field, "constant" coefficient form. Also, in practice one would naturally pick Γ_R



to be geometrically simple, e.g., a box, circle, or sphere. The method essentially assumes that the given differential equation holds in the region Ω_R, that the given boundary conditions hold on Γ, and that the truncated problem is closed by choosing an artificial boundary condition on the artificial boundary Γ_R. The particular choice of this artificial boundary condition is crucial to the design of an effective algorithm. The specific boundary conditions described below were first developed and analyzed in [2].

As a prototype acoustic problem, consider the following Helmholtz type problem. We have

$$\Delta u + k^2 u = f \quad \text{in} \quad \Omega \tag{1}$$

$$\frac{\partial u}{\partial n} = g \quad \text{on} \quad \Gamma. \tag{2}$$

In order to ensure that a unique solution exists, we must also impose a radiation condition, i.e., the Sommerfeld condition (in R^3)

$$-iku + \frac{\partial u}{\partial r} = 0\left(\frac{1}{r}\right) \quad \text{as} \quad r \to \infty, \tag{3}$$

which essentially allows only outgoing waves. In (1) k is the given frequency and f is a given source which we assume is of compact support or decays sufficiently rapidly as $r \to \infty$. In the near field, (1) may be replaced by a variable coefficient equation, while in the far field, (1) may be replaced by more complicated constant coefficient differential equations such as a reduced convected wave equation. The boundary condition (2) may be replaced by a Dirichlet boundary condition and/or mixed boundary conditions.

Now, it is known [3] that the solution of our problem (1)-(3) may be represented by the convergent expansion

$$u = \frac{e^{ikr}}{kr} \sum_{j=0}^{\infty} \frac{F_j(\theta,\varphi)}{(kr)^j} \tag{4}$$

where θ, φ represent spherical angles. One easily verifies that for u given by (4)

$$-iku + \frac{\partial u}{\partial r} = 0\left(\frac{1}{r^2}\right) \quad \text{as} \quad r \to \infty \tag{5}$$

so that (3) is certainly satisfied.

Now, our truncated problem is given by

$$\Delta u_m + k^2 u_m = f \quad \text{in} \quad \Omega_R \tag{6}$$

$$\frac{\partial u_m}{\partial n} = g \quad \text{on} \quad \Gamma \tag{7}$$

$$B_m u_m = 0 \quad \text{on} \quad \Gamma_R \tag{8}$$

where B_m is an operator to be defined below. The goal on the one hand, is to choose the operator B_m so that the difference $(u - u_m)$, i.e., the difference in the exact solution of the infinite domain problem and the exact solution of the truncated problem, is as small as possible. On the other hand, we wish to choose B_m so that the discretization of (6)-(8) is straightforward.

Our choice of boundary conditions is motivated by the representation (4). Indeed, consider the family of operators

$$B_m = \prod_{j=1}^{m} \left(\frac{\partial}{\partial r} - ik + \frac{2j-1}{r}\right) = \left(\frac{\partial}{\partial r} - ik + \frac{2m-1}{r}\right) B_{m-1}. \qquad (9)$$

It is easily verified by applying (9) to (4) that

$$B_m u\Big|_{r=R} = 0\left(1/R^{2m+1}\right). \qquad (10)$$

Thus we see that by requiring that (8) hold, we are making an error of $0(1/R^{2m+1})$ at the artificial boundary Γ_R if the latter contains the sphere of radius R. Thus, insofar as minimizing the error made at the artificial boundary, one would like to choose m as large as possible. However, practical implementation considerations preclude choosing $m > 2$. Basically, the difficulty is due to the fact that B_m involves an m-th order partial differential operator with respect to r. For $m = 1$, this poses no problems; for $m > 1$, one can use the differential equation to eliminate all r-derivatives of order greater than one. However, this procedure proves to be practical only for the case $m = 2$. See [2] for details. For these reasons, we focus attention on the operators

$$B_1 = \left(\frac{\partial}{\partial r} - ik + \frac{1}{r}\right) \qquad (11)$$

and

$$B_2 = \left(\frac{\partial}{\partial r} - ik + \frac{3}{r}\right)\left(\frac{\partial}{\partial r} - ik + \frac{1}{r}\right). \qquad (12)$$

Note that the Sommerfeld operator is not a member of our family B_m.

We should monitor the error in the truncated problem as a function of the frequency as well; indeed, the constants in the order relation (10) are functions of k. The representation (4) suggests that the proper parameter is (kr), although we should keep in mind that the functions $F_j(\theta,\varphi)$ also depend on k. However, we have computational and theoretical evidence [2] that indeed as k increases, the error in our boundary condition decreases. This partially mitigates the resolution problem. It is well known that as k increases, one must decrease the grid size h in order to accurately resolve waves, (kh) being the relevant parameter. However, as k increases, we are able to bring the outer boundary closer to the inner boundary, thus making the computational region smaller.

In two dimensions, an analogous family of boundary operators may be developed. They are based on the asymptotic expansion

$$u \sim \sqrt{\frac{2}{\pi kr}}\, e^{i(kr-\pi/2)} \sum_{j=0}^{\infty} \frac{f_j(\theta)}{r^j} \,. \qquad (13)$$

(Actually, an exact representation, analogous to (4) is known, but as it involves Hankel functions, (13) is easier to work with.) The family of operators B_m is now given by

$$B_m = \prod_{j=1}^{m} \left(\frac{\partial}{\partial r} + \frac{4j-3}{2r} - ik\right). \qquad (14)$$

Also, for Poisson problems where $k = 0$ in (1), we may use the multiple expansion

$$u = \frac{1}{r} \sum_{k=0}^{\infty} \frac{F_j(\theta, \varphi)}{r^j}$$

(in R^3) to generate appropriate boundary operators. These simply turn out to be given by (9) with $k = 0$. Similarly, we may relate the Poisson problem in R^2 to (14) with $k = 0$.

In [2] are collected many computational and theoretical results concerning the accuracy of the solution of the truncated problem (6)-(8). Note that from (1), (2), (6)-(8), and (10) that the difference $\varepsilon = u - u_m$ satisfies the problem

$$\Delta\varepsilon + k^2\varepsilon = 0 \quad \text{in} \quad \Omega_R$$

$$\frac{\partial\varepsilon}{\partial n} = 0 \quad \text{on} \quad \Gamma$$

$$B_m\varepsilon = B_m u = 0\left(\frac{1}{R^{2m+1}}\right) \quad \text{on} \quad \Gamma_R .$$

The questions answered in [2] are how does the known error committed at the boundary Γ_R affect the error in the interior Ω_R and at the inner boundary Γ. Roughly speaking it is found that error in the interior is of $0(1/R^2)$ for the boundary condition B_1 and is of $0(1/R^3)$ for the boundary condition B_2. Also in [2] is found a discussion of the implementation of the boundary conditions B_1 and B_2 in conjunction with finite element discretizations of (6)-(8).

Although the results of [2] are for second order elliptic partial differential equations, the artificial boundary conditions described above have been successfully used in other settings, e.g., transonic flow calculations with a subsonic free stream [4].

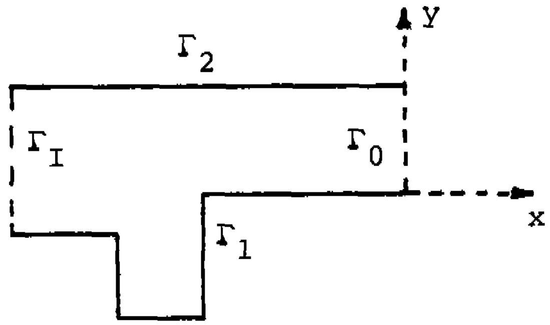
Recently, artificial boundary conditions have been studied for incompressible viscous flows in exterior domains [5]. Here, the governing equations are the Navier-Stokes equations. We assume that in the far field the flow approaches a uniform flow. Indeed, it is known that the velocity $\underline{u}$ and pressure p approach their free stream values $\underline{U}_{\infty}$ and p_{∞} at the rates $0(1/R)$ and $0(1/R^2)$, respectively. The following artificial boundary conditions have been studied. First, we simply impose $\underline{u} = \underline{U}_{\infty}$ at the artificial boundary. This is often done in the literature; however, we show that this is unsatisfactory from a computational point of view. Indeed, the $0(1/R)$ error made at the boundary Γ_R can lead to $0(1)$ errors in the interiors, and at best, in special situations, yields an $0(1/R^{1/2})$ error in the interior. A second boundary condition is to impose a zero stress boundary condition i.e.,

$$(p - p_{\infty})\underline{n} + \text{grad}\,\underline{u} \cdot \underline{n} = 0 \quad \text{on} \quad \Gamma_R. \tag{15}$$

We show that this boundary condition has an error of $0(1/R^2)$, i.e., the actual stresses on Γ_R are of $0(1/R^2)$. Furthermore, we show that the solution of the truncated problem then differs from that of the infinite problem by $0(1/R)$ in the interior of the truncated region. Although this is a significant improvement over the errors incurred by specifying the velocity on Γ_R, we feel that the results are still not good enough for practical computations. Indeed, we note that (15) plays the role of the Sommerfeld condition, i.e., they both are in an error of $0(1/R^2)$. Therefore, present work is centered on developing boundary conditions of higher accuracy, analogous to the family B_m described above. The starting point, analogous to the representation (4), is the theory of hydrodynamic potentials.

To close this section, we make some remarks concerning finite element discretizations of the truncated problem. Once we have defined our truncated problem, e.g., (6)-(8), it seems like a straightforward task to discretize it by a finite element (or finite difference) scheme. After all, the truncated problem has the appearance of a standard bounded domain problem. However, from a practical point of view, a naive discretization of a problem such as (6)-(8), e.g., using quasi-uniform grids, will usually require too many degrees of freedom, and thus lead to an inefficient algorithm. Advantage must be taken of the fact that the exact solutions decay as r increases. The way to do this is, of course, to grade the mesh, using increasing mesh sizes as r increases. The apparent decrease in accuracy caused by the increasing mesh size is counteracted by the fact that the exact solution, measured in the norms appearing in the error estimate, is decaying. By balancing these effects, one can easily design near optimal grids which lead to great savings in computer memory and processing time when compared to a naive approach. See [5] for a more detailed discussion of these ideas.

III. CHANNEL PROBLEMS FOR VISCOUS FLOWS. In this section we briefly consider downstream boundary conditions for incompressible viscous flows in channels. A typical configuration is given by the accompanying figure. Here, the boundaries Γ_1 and Γ_2 represent solid walls, Γ_I represents an inflow boundary, and Γ_0 represents an outflow boundary. Along Γ_I, Γ_1 and Γ_2, boundary conditions are given, e.g., along Γ_1 and Γ_2 the velocity vanishes and along Γ_I the velocity is specified. However, along Γ_0 a boundary condition is not known and therefore an artificial numerical boundary condition must be imposed in order to define a solvable discrete problem. In the absence of accurate asymptotic information at the outflow, one variously imposes some derivative of the solution at the outflow boundary. In [6], we carried out an analytical and computational study of the errors incurred by the imposition of such artificial boundary conditions. These studies were carried out for model linear transport equations and for the Navier-Stokes equations. Four choices for the artificial boundary conditions were considered. These involve specifying



$$u\Big|_{\Gamma_R} \tag{16}$$

or

$$\frac{\partial u}{\partial x}\Big|_{\Gamma_R} \tag{17}$$

or

$$\frac{\partial^2 u}{\partial x^2}\Big|_{\Gamma_R} \tag{18}$$

or

$$\int_{-d}^{0}\int_{0}^{y} u\,dx\,dy \tag{19}$$

where in the last integral, we refer to the coordinate system defined in the figure, and d is a fixed number. The first three choices seem "natural", and especially the second and third have been variously recommended in the literature. The last choice, although seemingly "unnatural", is of use in certain geophysical problems. See [6] for details. In (16)-(19), u represents a component of the velocity field, or perhaps the stream function or vorticity if a formulation involving those variables is employed. The actual values assigned to any of (16)-(19) are either determined by known boundary data (which of course, should be used whenever available), or is arbitrarily set, e.g., to zero, when unknown. The study of [6] was aimed at determining the effect of the latter type specification on the solution of the channel problem.

It was found that the incorrect specification of any of the boundary conditions (16)-(19) resulted in errors only in a boundary layer adjacent to the outflow boundary Γ_R. The thickness of this boundary layer decreased with increasing Reynolds numbers. See [6] for a detailed discussion.

ACKNOWLEDGEMENT. This report was prepared with the support of the Army Research Office under Contract No. DAAG 29-83-K-0084.

REFERENCES

1. M. Gunzburger and W. Layton, "On numerical boundary conditions for hyperbolic systems", Proc. of the 1981 Army Numerical Analysis and Computers Conf., ARO Report 81-3, 1981, 221-232.

2. A. Bayliss, M. Gunzburger and E. Turkel, "Boundary conditions for the numerical solution of elliptic equations in exterior regions", SIAM J. Appl. Math. 42, 1982, 430-451.

3. C. Wilcox,"A generalization of theorems of Rellich and Atkinson", Proc. Amer. Math. Soc., 1955, 271-276.

4. C. Cox, G. Fix and M. Gunzburger,"A least squares finite element scheme for transonic glow around harmonically oscillating airfoils", J. Comput. Phys., to appear.

5. G. Guirguis, "On the Stokes problem in exterior domains in R^3", Ph.D. thesis, University of Tennessee, 1983.

6. G. Fix and M. Gunzburger, "Downstream boundary conditions for viscous flow problems", Comput. Math. Appl. 3, 1977, 53-63.

LEAST SQUARES APPROXIMATIONS TO COMPRESSIBLE FLOW PROBLEMS

George J. Fix and Max D. Gunzburger
Department of Mathematics
Carnegie-Mellon University
Pittsburgh, PA 15213

ABSTRACT. A type insensitive scheme based on a weighted least square variational principle is used in conjunction with appropriate finite element spaces. Shocks are treated through a modified density formulation natural to this scheme. Selected numerical results are reported along with a qualitative analysis of the approximations.

I. INTRODUCTION.

Finite element schemes based on variational principles of the least squares type have a number of advantages as well as disadvantages when applied to compressible flow problems. The most striking advantage is their insensitivity to type; i.e., the approximations are typically just as accurate for supersonic flows as they are in the subsonic case. Moreover, in transonic flows, points in the supersonic region do not require special treatment. A closely related property is ability of this type of approach to generate Hermitian definite algebraic systems.

The primary defect of the least squares approach is that their accuracy deteriorates in the presence of shocks or other types of singularities. Moreover, this deterioration is far more severe than is seen in other types of finite element or finite difference formulations.

In this paper we treat this problem using suitably weighted norms in the least squares formulation. This approach does require a priori knowledge about certain qualitative features of the solution, but given this knowledge, it along with appropriate mesh refinement (or singular elements) can be used to accurately model singular behavior.

II. WEIGHTED LEAST SQUARES FORMULATIONS.

For simplicity we consider steady potential flows. The governing equation is

$$\text{div } \rho \, \overrightarrow{\text{grad}} \, \varphi = 0 \quad \text{in} \quad \Omega, \tag{1}$$

where φ is the velocity potential and ρ is the density. Bernoulli's equation permits us to write ρ as a function of the flow velocity $\vec{v} = \overrightarrow{\text{grad}} \, \varphi$. The problem statement is completed by supplying conditions on the boundary Γ of the flow region Ω:

$$\overrightarrow{\text{grad}} \, \varphi \cdot \underline{\nu} = v_\nu \quad \text{on} \quad \Gamma. \tag{2}$$

Here $\underline{\nu}$ is the outer normal to Γ and v_ν is the given normal velocity. On solid walls where there is no normal flow we have $v_\nu = 0$.

We reformulate (1) in terms of the mass flow

$$\vec{u} = \rho \, \overrightarrow{\text{grad}} \, \varphi. \tag{3}$$

Indeed, (1) is equivalent to

$$\operatorname{div} \vec{u} = 0 \tag{4}$$

$$\operatorname{curl}\left(\frac{\vec{u}}{\rho}\right) = 0 \tag{5}$$

where the density ρ is now regarded as a function of $\vec{u}$. In addition we assume the boundary condition takes the form

$$\vec{u} \cdot \underline{\nu} = u_{\nu} \quad \text{on} \quad \Gamma. \tag{6}$$

for a given normal mass flow u_{ν}.

To approximate this problem we select a finite element space $\mathcal{U}_h$, functions in which must be continuous. We then seek an approximate mass flow $\vec{u}_h$ in $\mathcal{U}_h$. This function is defined by first applying Newton's method to (4)-(6). Indeed, suppose $\vec{u}_h^{(0)}$ is an initial guess to the mass flow, and let

$$\sigma(\vec{u}) = \frac{1}{\rho(\vec{u})}. \tag{7}$$

Then one step of Newton's method produces

$$\vec{u}_h^{(1)} = \vec{u}_h^{(0)} + \vec{U}_h \tag{8}$$

where the correction $\vec{U}_h$ satisfies the linear problem

$$\operatorname{div} \vec{U}_h = -\operatorname{div} \vec{u}_h^{(0)} \tag{9}$$

$$\operatorname{curl}(\sigma_0 \vec{U}_h) + \operatorname{curl}[(\operatorname{grad} \sigma_0 \cdot \vec{U}_h)\vec{u}_h^{(0)}] = -\operatorname{curl}(\sigma_0 \vec{u}_h^{(0)}) \tag{10}$$

$$\vec{U}_h \cdot \underline{\nu} = -\vec{u}_h^{(0)} \cdot \underline{\nu} \quad \text{on} \quad \Gamma. \tag{11}$$

In (10), σ_0 denotes $\sigma(\vec{u}_h^{(0)})$. To solve (9)-(10) we use the least squares method. In particular, we require that $\vec{U}_h$ minimizes

$$\int_{\Omega} \{|\operatorname{div}[\vec{U} + \vec{u}_h^{(0)}]|^2 w_1 + |\operatorname{curl}[\sigma_0 \vec{u}^{(0)} + \sigma_0 \vec{U} + (\operatorname{grad} \sigma_0 \cdot \vec{U})\vec{u}^{(0)}]|^2 w_2\} \tag{12}$$

as $\vec{U}$ ranges over all functions in $\vec{\mathcal{U}}^h$ satisfying the boundary condition (11). In (12), w_1 and w_2 are weighting functions which vary with spatial locations.

The iterations are continued until the corrections $\vec{U}_h$ are suitably small. One attractive feature of this approach is that typically few Newton iterations are required. In addition, the least squares formulations for the corrections $\vec{U}$ lead to positive definite Hermitian systems which can be readily solved by a number of methods ([1],[2]). Finally, the approach is not sensitive to the presence of supersonic region where the equations are hyperbolic.

The approach does, however, have a major drawback. At points where the solution $\underline{u}$ is singular, or in the case of shocks, where it has very large gradients, there is a serious loss of accuracy in the approximation. This is where the weighting functions w_1, w_2 play a major role.

First suppose the flow region has a re-entrant corner as is shown in Figure 1. If the flow is subsonic at the corner A, then it is known that $\underline{u}$

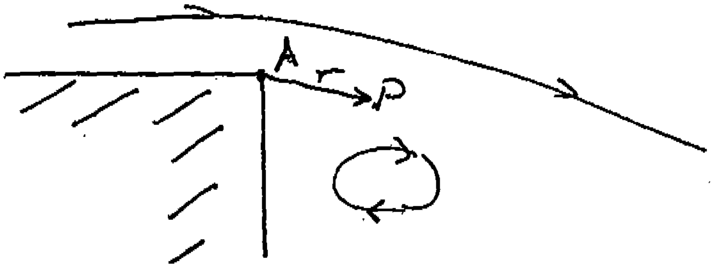


Figure 1. Corner singularity

behaves like $O(r^{-\alpha})$ for a suitable number $\alpha > 0$. In this case if the weights w_1, w_2 vanish at A like $O(r^{\beta})$, where $\beta > 2\alpha$, then with appropriate mesh refinement full accuracy can be achieved. Numerical experiments [3] indicate that the correct weights are essential. Mesh refinement alone is not sufficient.

A similar situation exists with shocks. Here it has been observed that it is sufficient for the weights to vanish like the first power of the distance to the shock. It is also necessary to add a small amount of dissipation, either through the density ρ or by a dash pot term.

III. A NUMERICAL EXAMPLE. To illustrate the above ideas we consider a time periodic flow over an oscillating plate as shown in Figure 2. Boundary conditions and other details are given in [3]. An exact solution can be

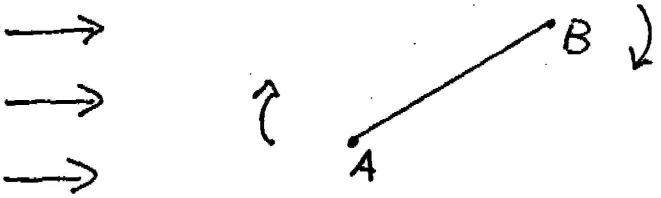


Figure 2. Oscillating plate

obtained in the case of a small time periodic disturbance of a uniform flow, and we compare this with the approximations obtained from the least squares method. The velocity field (and hence the mass flow) has a singularity like $O(r^{-1/2})$ at the tip of the plate (point A in Figure 2).

The pressure coefficient ΔC_p (which is proportional to the mass flow) is plotted in Figure 3 as a function of the distance along the plate. The solid line is the exact solution. The circles represent the weighted least squares approximation with mesh refinement, while the squares represent the unweighted least squares approximation on the same grid. The serious loss of accuracy in the latter is evident as the singularity at A is approached. Moreover, additional mesh refinement only makes marginal improvements. However, by adding the weights, which in this case behave like O(r), there is a significant improvement in the accuracy.

Additional examples are given in [3] including those containing shocks.

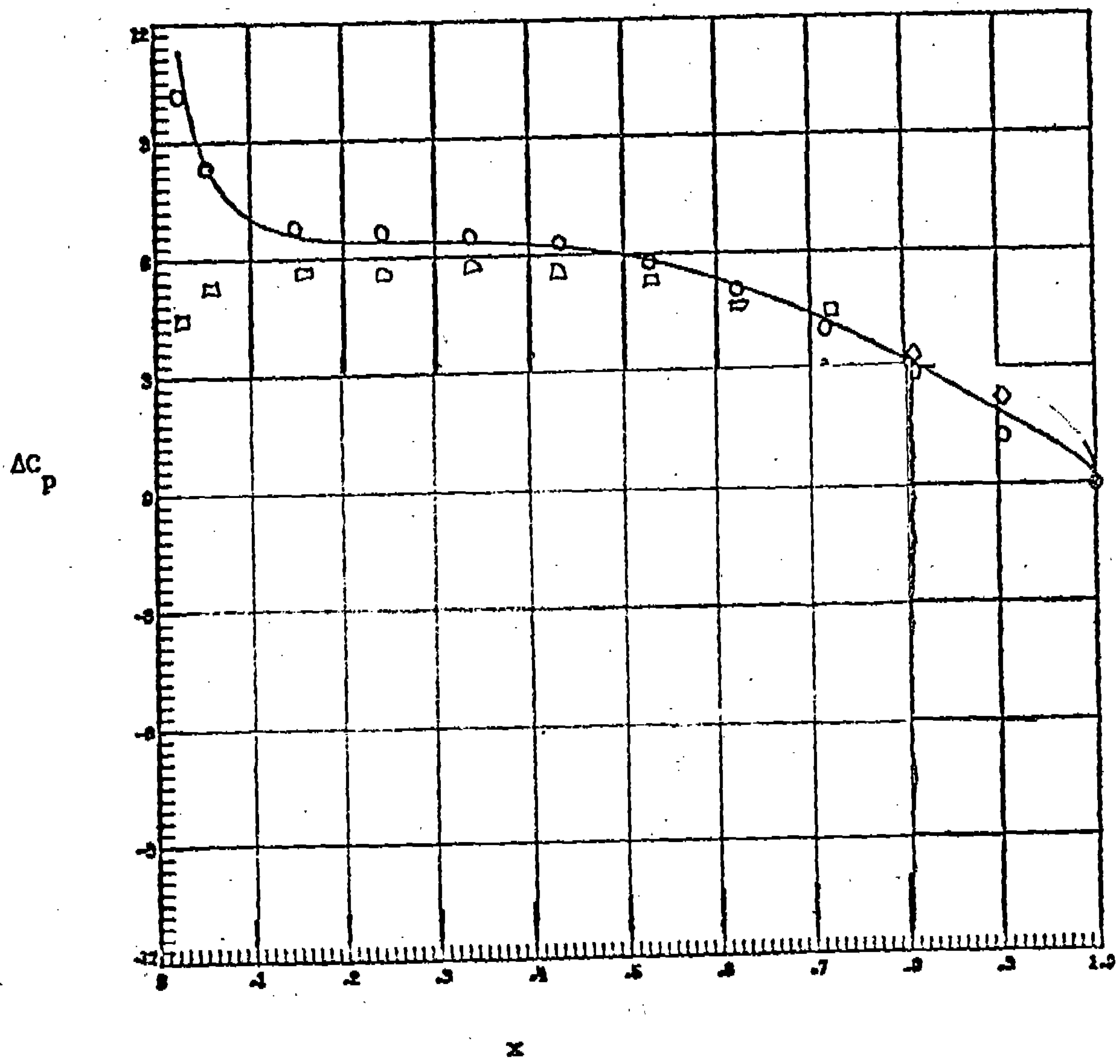


Figure 3. Pressure coefficient along plate

REFERENCES

[1] G. J. Fix and M. D. Gunzburger, "On least squares approximation to indefinite problems of the mixed types," Int'l. Journal for Numer. Mtd. in Eng., 12, (1978), pp. 453-470.

[2] G. J. Fix, M. D. Gunzburger, and R. A. Nicolaides, "Least squares finite element methods," NASA-ICASE Report 77-18, revised version published in Math. and Comp. with Appls., 5, (1979), pp. 87-98.

[3] C. L. Cox, G. J. Fix, and M. D. Gunzburger, "A least squares finite element scheme for transonic flow around harmonically oscillating wings," Journal of Comp. Physics, to appear, (1983).

AN INTRODUCTION TO GEOMETRIC PROGRAMMING

Patrick D. Allen
U.S. Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814
A.V. 295-5236

and

David W. Baker
Getty Oil Company Rm 1900
3810 Wilshire Blvd.
Los Angeles, CA 90010
(213) 739-2763

ABSTRACT. Geometric Programming (GP) can solve two types of problems: non-linear minimization formulations and simultaneous non-linear equations. Since a large number of problems in this world are non-linear, there are numerous potential applications. Furthermore, GP can provide the general solution to a given non-linear problem if the problem is reduced to what is called zero degrees of difficulty. Sample solutions to some problems are provided.

I. INTRODUCTION. Geometric Programming is mathematically simple to use. One does not have to take derivatives, pivot, or transform into some exotic N-space. All that is required is high-school level algebra and a bit of practice. One doesn't even need a computer, although for larger problems it can be useful. Many small problems can be solved by hand for the general case without even using a calculator.

If this sounds like a sales pitch, it is. Geometric Programming can be a nice alternative to trying to stuff a non-linear problem into a piecewise-linear Linear Programming package, or using a huge Conjugate Direction package on a relatively small non-linear problem. Furthermore, GP doesn't run forever. When compared to several different non-linear techniques, Geometric Programming has tended to be faster.

This is not to imply, however, that Geometric Programming is the cure-all of every non-linear problem, and it generally doesn't do well against linear problems. The restrictions on the uses of GP are presented below, as well as what types of problems Geometric Programming really prefers.

II. Preferences of Geometric Programming. As mentioned above, the two types of problems which GP can solve are:

1) non-linear minimization optimization, and
2) solutions to simultaneous non-linear equations.

Only minimization optimization problems can be solved by GP, but of course one can always transform a maximization problem into a minimization problem by suitably transforming the objective function. This transformation can be done by either multiplying the objective function by minus one, or by inverting the objective function (minimizing the reciprocal).

GP also prefers non-zero solutions. Since many non-linear problems have multiple solution points, especially in simultaneous equation problems, one does not know a-priori which solution will be found. However, one can be guaranteed that if there is any other solution than the zero solution, GP will find it. This factor can sometimes be used as an advantage when one desires to find multiple solution points.

Another feature of GP is that at optimality, all of the constraints will be "tight". In this context, "tight" means that all of the inequalities will be equalities. If this is not supposed to be the case at optimality, then the user must define his/her own artificial variables, just as is done in Linear Programming.

Geometric Programming also prefers that the problem is solved once in the general case rather than solved iteratively. (Don't we all!) This can be accomplished when the problem is formulated in the proper manner, called zero Degrees of Difficulty. This, too, will be explained later in more detail. The point we wish to make here is that sensitivity analysis is made trivially easy when the general solution to a problem is known, and GP allows the user to find the general solution without taking derivatives.

Problems involving trigonometric functions and powers to powers cannot yet be solved by GP. Also, there is no particular preference for integer solutions. Unlike the Transshipment problem in Linear Programming, even all-integer coefficients will not guarantee integer solutions in GP. There is no way, yet, to force integer solutions in Geometric Programming.

There are a couple of transformations which will allow the user to solve non-linear problems which include logarithms and exponential functions. Since these are two of the more common functions in non-linear problems, these transforms come in handy. We will not go into the details here, but they may be found in Beightler and Phillips.[1]

The types of problems which GP really likes are those with lots of variables multiplied together, with each variable to a positive or negative real power. For example:

$$\text{Minimize} \quad 2.45 * X^{2.2} * Y^{1.3} + 3.14 * X^{-3.1} + 4.02 * Y^{-.5} \quad (1)$$

A constant times a variable to a real power times another variables to a real power, plus another term of the same form, is the type of problem formulation Geometric Programming really prefers. Before explaining the solution to the above example, let us first define some terminology.

III. Terminology. The following definitions are used in Geometric Programming in order to simplify the discussions to follow.

A Term: Any group of variables separated by a plus or minus sign. In our example, there are three terms, one involving X, one involving Y, and one involving both X and Y.

An Objective Function: That which is to be minimized. In our example, it is all of equation (1).

A Constraint: That which must be satisfied while minimizing. There are no constraints in our example, but usually they are of the form of a sum of terms less than or equal to a constant.

A Weight or Delta: The contribution of that term to the final solution. In our example there would be three weights--one for each term. At optimality, the first delta would be equal to the contribution of the first term to the overall value of equation (1). This will turn out to be a number greater than zero and less than one.

Degrees of Difficulty: This is a measure of how difficult it is to solve the problem in its current formulation. This measure is defined to be equal to:

Number of Terms - Number of Variables - 1

This is similar to having the same number of equations as unknowns in order to solve for the variables. The Degrees of Difficulty are ranked by number.

Zero Degrees of Difficulty (0DD) is good. One can find the general solution to the problem if the sign of all the coefficients is positive. (If any are negative, you have to check to see if you are at a minimum, a maximum, or a saddle point.) If the Degree of Difficulty is one (1DD), one can find the general solution, but only after taking logs, derivatives, and exponentiating. This is messy and is not usually done unless the genral solution is required. Any higher number for Degrees of Difficulty requires that iterative solutions be used in order to find the optimal solution. Using iterative solutions does not guarantee a global optimal solution, but it usually works out that way. Furthermore, the number of iterations required by GP is small compared with most other search algorithms. It has been found that GP performs at least as well as Newton's Method applied to a log transform of the original problem, and is probably better.[2] Research continues on this subject.

IV. The Rules of Geometric Programming. There are four rules in Geometric Programming which result from the theory behind GP. These shorthand rules allow the user to solve properly formulated GP problems. They are listed below:

1) $$\phi^* = \prod_{i=1}^{N+\ell m} \left(\frac{k_i}{\delta_i}\right)^{\delta_i} \prod_{j=1}^{m} \left(\left(\sum_{k=1}^{\ell} \delta_{jk}\right)^{\left(\sum_{k=1}^{\ell} \delta_{jk}\right)} \right) \text{FOR } j \text{ CONSTRAINTS}$$

2) $$\delta_1 + \delta_2 + \ldots + \delta_N = 1$$

3) $$\phi^* = \frac{\text{1ST TERM}}{\delta_1} = \frac{\text{2ND TERM}}{\delta_2} = \ldots = \frac{\text{NTH TERM}}{\delta_N}$$

4) $$\delta_{jk} = \left(\text{JKTH TERM}\right) \sum_{k=1}^{\ell} \delta_{jk}$$

Rule 1 is the value of the objective function at optimality, represented by Ø* (phi star). The K's are the coefficients if each of the terms in the whole problem, the Deltas are the weights for each term. The superscripts and subscripts show that there are N terms in the objective function, and ℓ terms in each of the m constraints. The deltas in the constraints have two subscripts to represent the kth term in the jth constraint. We shall show in our examples how to solve for the weights or deltas. Once these deltas are known, then the optimal solution is known by solving for phi star in Rule 1. This means that the final solution is known prior to the values for the variables are determined. Again, no pivoting is required.

Rule 2 restates the percentage rule--each delta of the objective function will, at optimality, be equal to the contribution of the term with which it is associated. If the first term contributes 50% of the value of equation (1), for example, then the first delta will equal one half. Since the sum of the parts must equal the whole, the sum of the deltas of the objective function must equal one. This rule also provides an extra equation with which to solve for the deltas, and is responsible for the negative one in the definition of Degrees of Difficulty.

Rule 3 Is a restatement of the definition of the deltas--each delta represents the contribution of its term to the whole objective function. Rule 3 is used to determine the optimal values of the variables after the optimal solution has been derived. In our example, the variable X may be determined by using Rule 3 applied to the second term. Since all values are known except for X, one simply solves a single term monomial to determine X.

Rule 4 is similar to Rule 3, except that Rule 4 holds for deltas in the constraints rather than in the objective function. The contribution of the kth term in the jth constraint is equal to its associated weight divided by the sum of all of the weights in the constraint. This is used whenever it is more difficult to solve for the variables using Rule 3 than it is to use Rule 4.

V. Solving For the Weights or Deltas: Now that we have determined terminology and how to solve the problem given the appropriate weights, let us now examine how to solve for those weights. In our example, we have three terms and two variables, so we have a zero Degree of Difficulty problem. Furthermore, by Rule 2, we know that the sum of the deltas must equal one. This gives us our first equation:

$$\delta_1 + \delta_2 + \delta_3 = 1 \tag{2}$$

Notice that since the deltas are the unknowns, we will need two more equations in order to solve for the deltas explicitly.

Next, we look to see in which terms the variable X appears. If X does not appear in the term, then the coefficient for the delta in our next equation is zero. If X does appear in the term, then the coefficient of the delta is equal to the exponent of the variable X. In our example,

$$2.2 * \delta_1 - 3.1 * \delta_2 = 0 \tag{3}$$

and the right hand side is defined to be equal to zero.

The same thing is done for Y as was done for X. In our example,

$$1.3 * \delta_1 - .5 * \delta_3 = 0 \tag{4}$$

which completes the number of required equations with which to solve for the deltas. Notice that even though the problem as originally defined was non-linear, the solution to the problem of finding the weights for each term is a linear problem. Furthermore, no derivatives had to be taken and no pivoting was required.

The values of the deltas in the above problem are:

$$\delta_1 = .232036, \quad \delta_2 = .16467, \quad \text{and} \quad \delta_3 = .60329$$

From Rule 1 we find that the optimal solution equals 8.81597, and by applying Rule 3 to the second and third terms, X equals 1.28256 and Y = .57129. All that was used to solve this problem was a little paper and a hand calculator. The solution was verified by substituting the optimal values for X and Y back into the original equation, and seeing if that number matched the solution given by GP. One can also assure that the solution is a minimum by changing the value of X or Y slightly in equation (1) and showing that the result is an increase in the value of the objective function.

VI. An Example Comparing Lagrange Method to GP: In the following example, we will first solve the problem by means of the Lagrange Method which requires the taking of three partial derivatives to find the optimal value. Then we will show how the same solution can be accomplished through Geometric Programming.

The problem is to minimize the amount, and therefore the cost, of material required to build a cylindrical storage tank which is constrained to contained at least 1000π cubic yards of liquid. The pi is obviously there to make the solution come out as nice round numbers. The problem statement and results are as follows:

$$\text{MIN } \$ = \$\pi R^2 + \$2\pi RH$$

$$\text{S.T.} \quad \pi R^2 H \geq 1000\pi \text{ CUBIC YARDS}$$

$$L(H,R,\lambda) = F(H,R) - \lambda(G(H,R) - K_1)$$

$$R^3 = 1000 \quad R^* = 10 \quad H^* = 10$$

$$\text{COST} = \pi R^2 + 2\pi RH = \$300\pi = \$942$$

In order to find the solution, the Lagrange (L(H,R, λ)) was required to have the partial derivative taken with respect to R, H, and λ . These partials were set equal to zero, and the resulting three equations and three unknowns (R, H, λ) were solved. The results were that the radius and the height of the cylinder (with no bottom--the bottom is a sunk cost) are both equal to 10 yards. The total cost of the operation was $942.

In order to solve the problem by Geometric Programming, one must first get the problem in the proper form. All of the constraints must be of the form where all of the terms are less than or equal to one, as shown below.

$$\text{MIN } \$ = \$\pi R^2 + \$2\pi RH$$

$$\text{S.T. } 1 \geq \frac{1000\pi}{\pi} R^{-2} H^{-1}$$

Once in this form, ensure that the Degree of Difficulty is zero. If so, then set up the delta matrix, solve for the deltas, determine the optimal solution through Rule 1, and find the values for R and H using Rule 3 and/or Rule 4.

$$\text{MIN } \$ = \$\pi R^2 + \$2\pi RH$$

$$\text{S.T. } 1 \geq \frac{1000\pi}{\pi} R^{-2} H^{-1}$$

$$\delta_1 + \delta_2 = 1$$

$$R: \quad 2\delta_1 + \delta_2 - 2\delta_3 = 0$$

$$H: \quad 0\delta_1 + \delta_2 - \delta_3 = 0$$

$$\delta_1 = 1/3 \qquad \delta_2 = 2/3 \qquad \delta_3 = 2/3$$

$$\phi^* = (3\pi)^{1/3}(3\pi)^{2/3}(1000)^{2/3} = 3\pi 100 = \$942$$

VII. For Simultaneous Equation Problems: The next example will illustrate how to solve for two variables in two non-linear equations. We will use the method created by Baker[3], which ensures that all of the coefficients will be positive. The method created by Allen[4] allows for negative coefficients, and sometimes finds a different, though equally valid solution point. The two methods work well together, so that problems in which one has difficulty solving, the other has no problem at all.[5]

The problem is stated as:

$$x_1^2 + x_1x_2^3 = 9$$

$$3x_1^2x_2 - x_2^3 = 4$$

which is placed in proper format by solving for the right hand side to be equal to one.

$$\frac{1}{9}x_1^2 + \frac{1}{9}x_1x_2^3 = 1$$

$$\frac{3x_1^2 x_2}{4 + x_2^3} = 1$$

The calculations are provided in Appendix A. In summary, one first performs the operation known as "Condensing" on the two equations, and seperately on the numerator and denominator if the equation is a ratio. To condense, one simply chooses a value for each of the two variables (guess a solution) and determine how much each of the variables contribute to each equation at these chosen values. These are used as weights in order to solve for the variables. These new values for the variables are then substituted back into the original equations, and new weights are determined. These weights, in turn, are used to solve for new values for the variables and the pattern keeps repeating until the change from one iteration to the next is as small as desired.

The advantage is that the original guess by the user does not have to be accurate at all. One can choose a value of 1000 for the first variable, and in one iteration, the value is of the order of magnitude (one) of the optimal solution. One solution to this problem is:

$$x_1 = 1.33636$$

$$x_2 = 1.75424$$

<u>VIII. A Final Example:</u> As a final example, we would like to present a problem presented to us by an analyst at the Concepts Analysis Agency. The problem was to create an allocation scheme by which aircraft could be allocated to either near or far targets to best reduce the enemy's capabilities. The key restriction was that whatever scheme was used should be flexible enough to adapt to anyone's rational idea of how they would perform the allocation scheme. This was necessary since not only would analysts differ as to which scheme should be used, but also the commanders who were being modelled would also be likely to differ in their allocation schemes. Therefore, a truly flexible scheme was required.

The following formulation was proposed. Let K_1 be the enemy force near to friendly forces, and let K_2 represent the forces farther away. The second number could be discounted over time or distance or both. Let N_1 represent the number of aircraft allocated to near targets, and N_2 represent the number of aircraft allocated to far targets. The sum of the aircraft must be less than or equal to the total number of aircraft available, N. The objective is to minimize the total amount of (discounted) force facing the friendly forces, and this is represented by decreasing the enemy force by the number of aircraft allocated to some power $-\alpha$. We assume for a moment that enemy air defense threat is the same, that α is the same for both near and far targets, and no aircraft need be held in reserve.

The formulation and solution are given below:

$$\text{MIN} \quad K_1 N_1^{-\alpha} + K_2 N_2^{-\alpha}$$

$$\text{S.T.} \quad N_1 + N_2 \leq N$$

$$\text{LET } B = \left(\frac{K_2}{K_1}\right)^{\frac{1}{1+\alpha}}$$

$$N_1^* = \frac{N}{1+B} \qquad N_2^* = \frac{NB}{1+B}$$

Notice that since there are four terms and only two variables, this is a one Degree of Difficulty problem. As mentioned previously, logarithms and derivatives need to be taken. Once again, we will not go into the gory detail here. The important point is to examine the sensitivity of the solution and its formulation. Let alpha equal 0. This results in the standard "weighted" allocation scheme where the portion allocated to each target is same as the proportion of each target to the total threat.

If we were to let alpha equal 1, we would have a "modified" weighting scheme, where the portion allocated to each target is a function of the magnitude of each threat. Similarly for higher values of alpha. If, instead, we were to let alpha equal -1, then all aircraft would be allocated to whichever threat was larger. This is consistent with the fact that when alpha equals -1, the original problem is linear. The solution to a linear problem with two variables is always to find the "corner" solution point; that is, to allocate everything to the varaible with the highest payoff.

Finally, if we were to let alpha equal infinity, then half of the aircraft would go to the near target, and half to the far target, no matter what the values of the coefficients. These four results are summarized below.

$$\text{IF } \alpha = 0; \quad N_1 = N \frac{K_1}{K_1 + K_2}, \quad N_2 = N \frac{K_2}{K_1 + K_2}$$

$$\text{IF } \alpha = 1; \quad N_1 = N \frac{\sqrt{K_1}}{\sqrt{K_1} + \sqrt{K_2}}, \quad N_2 = \frac{\sqrt{K_1}}{\sqrt{K_1} + \sqrt{K_2}}$$

$$\text{IF } \alpha = -1; \quad N_1 = 0, \quad N_2 = N \quad \text{IF } K_2 > K_1$$

$$N_1 = N, \quad N_1 = 0 \quad \text{IF } K_1 > K_2$$

$$\text{IF } \alpha = \infty; \quad N_1 = \frac{N}{2}, \quad N_2 = \frac{N}{2}$$

The advantage to the above allocation scheme is that it is flexible enough to contain all of the "standard" allocation schemes used--a weighted scheme, all or nothing, or half and half. Furthermore, by having solved for the general solution, one can calculate the optimal allocation given alpha without having to re-derive the problem each time alpha is changed. This also allows alpha to change value during the operation of the simulation without large blocks of code being required for each scheme.

IX. Conclusions: "If the world were linear, Dantzig would be king." This quote is based upon the fact that Dantzig was chosen to compile what was known about Linear Programming into the classic book on the subject. Solutions to linear problems are relatively easy to obtain with numerous packages available. However, the world is not linear, and a large portion of the problems in this world do not fit well into linear programs. One can theoretically always turn a non-linear problem into a very, very large piecewise linear problem, but this is awkward. There exist a number of non-linear packages which occupy large sections of core when running, and sometimes run forever.

Geometric Programming provides a nice alternative to solving non-linear problems, especially ones which include a managable number of variables (say less than twenty) and whose exponents are not all integers. Larger problems are, of course, solvable by Geometric Programming, but one should then examine the GP computer packages available, or consider writing a program to solve that unique problem.

APPENDIX A: SAMPLE SOLUTION TO A SIMULTANEOUS NON-LINEAR EQUATION PROBLEM

The problem as given stated:

$$X_1^{\,2} + X_1X_2^{\,3} = 9$$

$$3X_1^{\,2}X_2 - X_2^{\,3} = 4$$

The condensed formulation of the problem is obtained as follows:

$$\frac{1}{9} X_1^{\,2} + \frac{1}{9} X_1X_2^{\,3} = 1$$

$$\frac{3X_1^{\,2}\ X_2}{4 + X_2^{\,3}} = 1$$

So condensing yields:

$$(\tfrac{1}{9})\ (\frac{X_1^{\,2}}{W_{11}})^{W_{11}}\ (\frac{X_1X_2^{\,3}}{W_{12}})^{W_{12}} = 1$$

$$\frac{3X_1^{\,2}X_2}{(\frac{4}{W_{21}})^{W_{21}}(\frac{X_2^{\,3}}{W_{22}})^{W_{22}}} = 1.$$

Taking logs and setting up the matrix form of the system gives:

$$\begin{bmatrix} (2W_{11} + W_{12}) & (3W_{12}) \\ (2) & (1-3W_{22}) \end{bmatrix} \begin{bmatrix} \ell n\ (X_1) \\ \ell n\ (X_2) \end{bmatrix} = \begin{bmatrix} -\ell n\ (\frac{1}{9}(\frac{1}{W_{11}})^{W_{11}}\ (\frac{1}{W_{12}})^{W_{12}}) \\ -\ell n\ (3(\frac{W_{21}}{4})^{W_{21}}\ (W_{22})^{W_{22}}) \end{bmatrix}$$

Using a starting point of $X_1 = 50$, $X_2 = 20$, and letting

$$K_1 = (\frac{1}{9})\ (\frac{1}{W_{11}})^{W_{11}}\ (\frac{1}{W_{12}})^{W_{12}}$$

and

$$K_2 = 3(\frac{W_{21}}{4})^{W_{21}}\ (W_{22})^{W_{22}}$$

results in:

$$W_{11} = \frac{X_1^{\ 2}}{X_1^{\ 2} + X_1X_2^{\ 3}} = \frac{2500}{2500 + 400000}$$

$$= 0.00621$$

$$W_{12} = \frac{X_1X_2^{\ 3}}{X_1^{\ 2} + X_1X_2^{\ 3}} = \frac{400000}{2500 + 400000}$$

$$= 0.99379$$

$$W_{21} = \frac{4}{4 + X_2^{\ 3}} = \frac{4}{4 + 8000}$$

$$= 0.00050$$

$$W_{22} = \frac{X_2^{\,3}}{4 + X_2^{\,3}} = \frac{8000}{4 + 8000}$$

$$= 0.99950$$

$$K_1 = (\tfrac{1}{9}) (\frac{1}{0.00621})^{0.00621} \quad (\frac{1}{0.99379})^{0.99379}$$

$$= 0.11538$$

$$K_2 = (3) (\frac{0.0005}{4})^{0.0005} \quad (\bar{0}.9995)^{0.9995}$$

$$= 2.98505$$

Solving the system [2.31] for $\ell n(X_1)$ and $\ell n(X_2)$ yields:

$$\ell n \; (X_1) = 0.26179$$

$$\ell n \; (X_2) = 0.68014$$

so

$$X_1 = 1.29925$$

$$X_2 = 1.97415.$$

The process is then repeated by calculating new weights and continuing as before. A summary of the iterations is provided below:

Iteration	X_1	X_2
0	50.0000	20.0000
1	1.2992	1.9742
2	1.3304	1.7645
3	1.3363	1.7543
4	1.3363	1.7542

LITERATURE CITED

(1) Beightler, Charles S., and Don T. Phillips. 1976. Applied Geometric Programming. New York: John Wiley & Sons, Inc., p. 291.

(2) Greening, Doran R. 1981. "A Proof of Convergence for a Condensation Approach to the Solution of Systems of Nonlinear Equations." Ph.D. Dissertation, Colorado School of Mines, p. 3.

(3) Baker, David W. 1980. "Application of the Geometric Inequality to the Solution of Systems of Nonlinear Equations." Ph.D. Dissertation, Colorado School of Mines, pp. 26-30.

(4) Allen, Patrick D. 1980. "Unbounded Weights: A New Approach to the Solution of Nonlinear Optimization Problems with Applications to Mineral Economics." Ph.D. Dissertation, Colorado School of Mines, pp. 29-31.

(5) Bernard, Paul L. 1982. "An Algorithm for the Solution of Investment Problems Having Dual Rates of Return." Ph.D. Dissertation, Colorado School of Mines, pp. 37-38.

SPACE AND TIME ANALYSIS IN DYNAMIC PROGRAMMING ALGORITHMS

NAZIR A. WARSI AND C. BENNETT SETZER

Department of Mathematical and Computer Sciences
Atlanta,University
Atlanta, Georgia 30314

The work upon which this report is based was supported by the U.S Army Research Office and the Atlanta University.

ABSTRACT

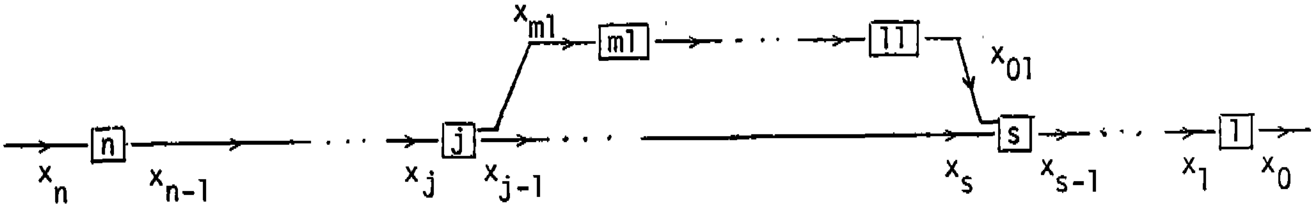
Dr. Nazir Warsi, in recent work, showed how to solve certain dynamic programming problems while keeping strict bounds on the amount of working storage needed. We discuss extensions of Dr. Warsi's methods and analysis to more general dynamic programming networks. We describe a general algorithm for solving problems in this more general class. This algorithm may be applied in such a way as to limit working storage arrays to any dimensions greater than or equal to 2. In making this restriction, there are two costs: a number of arrays of dimension 2 may need to be stored simultaneously; searches for maxima can become arbitrarily complex with the complexity of the network.

We discuss the implementation of the general algorithm in a higher level language with particular emphasis on storage management. We also discuss data representations and the practicality of implementing a system for handling general networks.

The storage and computational requirements of dynamic programming algorithms have limited their practical uses. The storage and computational demands can become excessive whenever the state vectors are of dimension 3 or more, or if the number of states grow exponentially. The purpose of the research described in this paper is to implement dynamic programming solutions for certain basic kinds of networks in a manner that will keep the working storage needs to a minimum with a possible trade-off in increased computing time.

The types of networks considered in this study are converging branch, diverging branch, feed-forward loop, and feed-back loop. In this note, we will present the implementation and analysis of alsorithms for the feed-forward loop system. For further information, refer to the reports of Warsi and Esogbue in the references.

Feedforward Loop System Diagram



Each node in a dynamic programming network represents a decision point. The decision is a finite valued variable, values taken to be integers 1 .. p_i for node i. Edges connecting nodes are directed indicating a flow from one node to another. The edges may carry one of a finite number of state values, taken to be integers from 1 .. k_i for edge i. For each node, functions are defined determining the output states (on edges leaving the node) and a return (a numeric value). The function values depend on input states and on the decision value for that node. The solution of a network consists of a choice of input state for each edge with external source and a choice of decision for each node. Further, these choices will maximize the sum of the returns for all nodes in the network over all possible such choices.

The output state functions and the return functions can be represented by matrices, or higher dimension analogues. We will use r_i to denote the return function (array) for node i and t_i or t'_i to denote the output functions (array) for node i. Note that t' will only be defined for nodes with two output edges.

The strategy for solving a network is to iteratively combine return functions for nodes into functions representing the total return possible from groups of nodes. The value of such a combined return function will depend only on state values of edges that enter the represented group of nodes from outside the group or that leave the represented group. In particular, the value will not depend on state values on edges between nodes within the group.

For the feed-forward loop network, a starting function is computed which represents the maximal return from node 1 given the input x_1. This function is combined with returns from nodes back to node s - 1. During all of this, the function can be represented by a one-dimensional array. Upon combining with the return from node s, the function depends on two arguments (x_{s-1} and x_{01}), and is represented by a two-dimensional array. Returns from nodes on the parallel portions of the network are then combined until node j is one-dimensional array. Further iterations produces a function representing the group of all nodes in the network, return depending on x_n. Maximizing over x_n produces the network maximal return.

Analysis of the work done to this point shows that the working storage needed is simply the representation of the current combined function. If k is the maximal number of states on any edge, this amount is proportional to k^2, since a two-dimensional array was needed. However, even though the maximal total return from the network is known, and the free input x_n is known, the decisions necessary at each node are not. Approaches to determining these decisions fall on a spectrum between these two extremes: store for each node, as its return is entered into the combined function, the functional dependance of the decision made on the input states to the group; repeat the optimization computation above retaining functional information about the decision in only the last node processed. The latter approach works since the optimization computation is carried out for one less node at each stage. The former approach requires a large amount of storage, depending on the number of nodes in the network, among other variables. The latter approach requires a large amount of time, but only the working storage needed for the optimization computation (proportional to k^2).

One of us (Warsi, loc. cit.) has developed a variation on the recomputation approach that cuts down the number of recomputations needed considerably. In this method, optimal decisions are stored as functions of the group input states if all such states are on edges to the node being entered into the combined function. Extra storage is not needed in this case, as the t_i (or t_i') matrix can be used by marking appropriate entries. For example, if there is one input edge and one output edge, and the combined function being computed depends only on the state value on the input edge, then the marking method is as follows. For each input state, mark that entry in the corresponding state row of t_i that is in the column of the optimal decision for that input state. Since the entries in the t matrix are no larger than k (the maximal value of any state), this marking can be done by adding k to the entry. Doing this, the recomputations can be cut down to four, regardless of the number of nodes in the network. Similar savings in time and space are achieved for the other network types by using this method of saving decisions with limited recomputation.

The most interesting problems that arise in actually implementing the algorithms developed for the non-serial networks is the representing of the arrays in storage. We wished to represent the network and r and t functions in a uniform way that could be operated on whatever the distribution of k_i and p_i. Although PASCAL, the language we chose for implementation, does allow dynamic allocation, array sizes must be specified at compile time. (This will vary for different versions of PASCAL). We, therefore, chose (influenced by our FORTRAN background) to use one large one-dimensional array to store all information in. One extra array is used to point to the beginning of the information for each node in the big array. Various selector and set functions can then be written to access the information. If the algorithms are moved to a different language (ADA, for example), the various arrays could then be allocated separately, using pointers to link information together. Due to the sequential nature of the processing, double links between nodes would suffice to move around as needed within the network.

We have made some investigation of more complex networks. Work done by Onukwuli (see reference) had quantified some of the problems in the main computation that arise from attempting to keep the combined function of low dimension. It also appears that, in non-serial networks, that there are certain key points at which combined functions can be saved for future reference that will cut down on recomputation time. We hope that this will lead to reasonably space efficient methods for solving very complex dynamic programming networks.

References

The following two entries are the two parts of the final report for Army Research Project DAAG 29-80-G-0010.

Esogbue, Augustine O., DYNAMIC PROGRAMMING ALGORITHMS AND ANALYSES FOR NONSERIAL NETWORKS.

Warsi, Nazir, DYNAMIC PROGRAMMING ALGORITHMS AND ANALYSES FOR NONSERIAL NETWORKS.

Onukwuli, Francis, AN INVESTIGATION OF DYNAMIC PROGRAMMING NETWORK ANALYSIS OF COMPLEX NONSERIAL SYSTEMS, Thesis, Atlanta University, 1983.

MATHEMATICAL ANALYSIS OF THE COUNTERFIRE DUEL: TANKS VS. ANTI-TANK MUNITIONS

Joseph V. Michalowicz
Tactical Nuclear Warfare Branch
Harry Diamond Laboratories, USA ERADCOM
Adelphi, MD 20783

ABSTRACT. A detailed, analytic model is developed to represent the duel between a ground laser designator (GLD) directing a sequence of laser-guided rounds against a platoon of target tanks which counterfire against the GLD. The model accurately portrays the complex interplay between the designator-on time, the rate of fire of the laser-guided rounds and the tank counterfire response time distributions. Also taken into account are the tank aiming errors and range estimating techniques, the level of GLD protection, flight times, designation modes and degree of coordination of the tank platoon.

One of the innovations of this model is the utilization of gamma distributions to represent tank counterfire response times. This permits the representation of the time-to-fire for any number of tank rounds within the designator-on time interval by convolution of the gamma density functions. Exact expressions which allow for all encounter outcomes are derived for computation of the expected number of tanks killed and probability of GLD kill.

This model was used to determine the probabilistic outcomes of the encounters at each stage of the force-on-force analysis in the definitive U.S. Army study on GLD survivability.

I. INTRODUCTION. The development of laser-guided missiles and projectiles has provided a new dimension to the battlefield environment for tanks. From a remote ground or airborne forward observer (FO) position, a laser designator operator illuminates a target tank with a directed laser beam (see fig. 1). The laser energy reflected from this spot on the target then enables the seeker to guide the missile/projectile to the target tank. If able to detect the presence of the laser designation, the tank (and its support units) may use various countermeasures (CM), such as taking evasive action, using smoke or chaff to disguise the tank's position, generating false target images to deceive the missile seeker, dazzling the FO with flashlamps, and/or directing counterfire against the FO, the missile launching platform, or the missile itself. This report presents a rigorous analysis of the Tank Counterfire Duel, in which laser-guided missiles are directed against a tank or tank platoon via a ground laser designator (GLD), and the tanks detect the laser radiation and counter by firing their main guns in an effort to destroy the GLD.

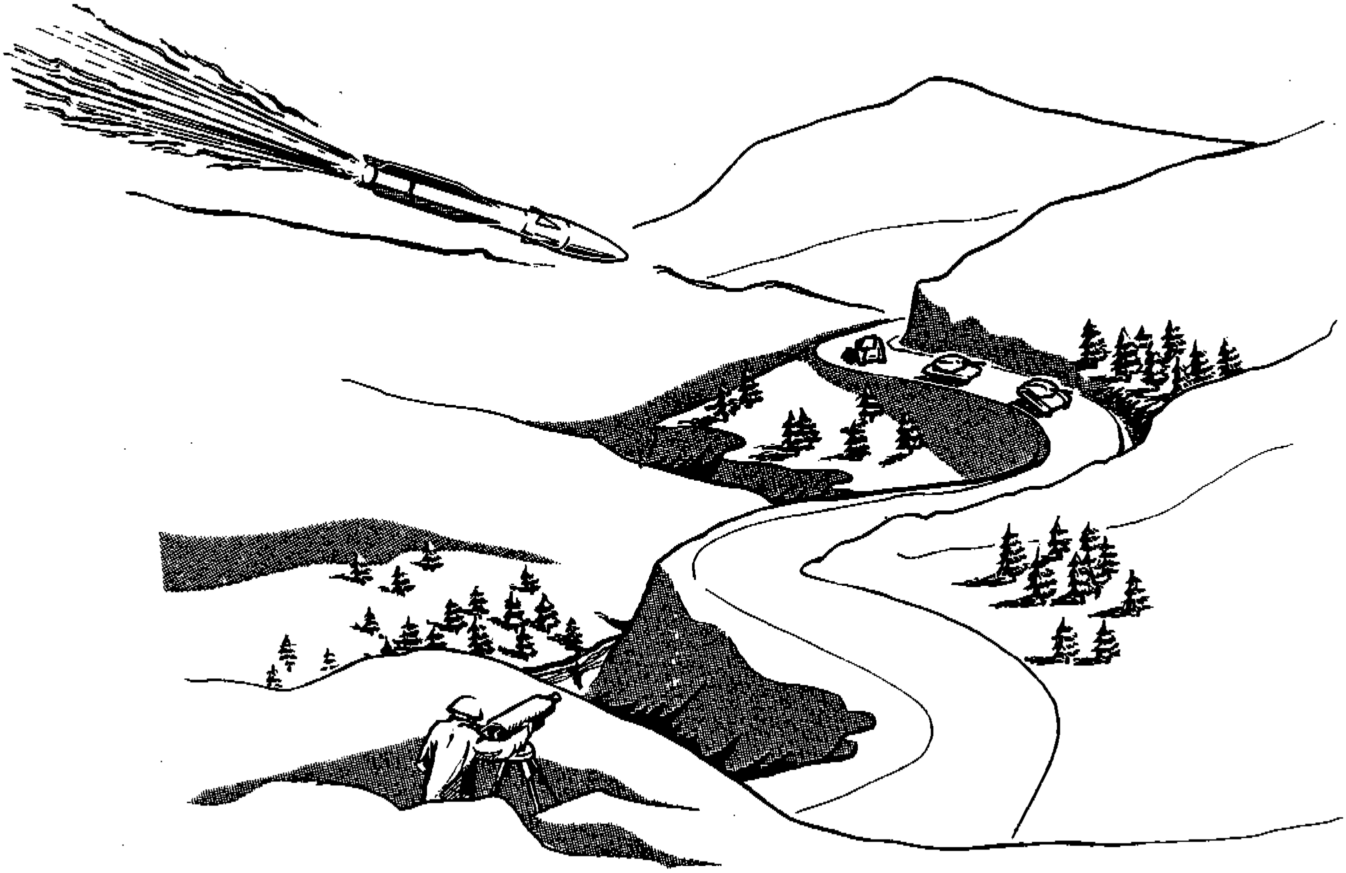


Figure 1. Laser-guided missile versus tank duel.

The analytic model which will be derived takes into account tank counterfire response time distributions, designator-on times, tank-to-GLD range, missile/projectile single-shot kill probabilities (SSKP), tank round SSKP against the GLD, tank fire control errors, GLD protection, ground slope at the GLD, missile and projectile flight times and rates of fire, tank round flight times, and degrees of coordination of the tank platoon. Special features which make this model unique include the following:

1. Probability distributions are derived to represent tank times-to-fire. In particular, gamma density functions are fit to the times-to-first-fire and time-between-fires data. The time-to-n^{th}-fire distribution is then obtained by convolution of the time-to-first-fire density with n - 1 copies of the time-between-fires density.
2. Any number of tank fires are allowed during the designator-on time interval.
3. Flight time of the tank round is included.
4. Various GLD positions and protection levels are considered.
5. Analytic expressions which incorporate the time-to-fire distributions and allow for all encounter outcomes are formulated for computation of the expected number of tanks killed and probability of GLD kill. This is not a simulation model!
6. In addition to the one round versus one tank duel, the encounter between three laser-guided rounds versus a platoon of three tanks is analyzed, with attention paid to the level of coordination of the tanks.

The data base used for the model was derived from the totality of U.S. Army field experiments on tank counterfire response and GLD kill. Most of these data cannot be detailed in this paper due to classification, but the form of the data will be discussed. GLD suppression, that is, degradation in GLD crew performance due to counterfire near-misses and obscuration, could not be included due to a lack of proper experimental data, even though the model could be adapted to handle suppression.

Because of Congressional concern over the survivability of ground laser designators on the battlefield, a special task force, the Survivability Study Task Force for Ground Laser Designators (SSTF), was established to answer the survivability question once and for all. The SSTF performed a two-stage analysis of the physical survivability of the GLD. First, the model described in this paper was used to provide a detailed analysis of the outcomes of counterfire duels played under a wide variety of conditions. These duels were then placed in a realistic battlefield context as the fixed-piece engagements occurring in a force-on-force map exercise pitting one Blue company on the defense against a Red tank threat. The resulting battlefield engagement assessment[1] provided a thorough and credible answer to the GLD survivability question, which has been cited frequently in Congressional testimony. In addition to the duel between one laser-guided round and one tank (1-on-1) and the encounter between three rounds fired in sequence and a platoon of three tanks (3-on-3) discussed in this paper, models for other encounter combinations were derived[2] for use in the SSTF map exercise.

[1] Thomas J. Gleason, Joseph V. Michalowicz, Morgan G. Smith, and Richard Scungio, Final Report -- Survivability Study Task Force for Ground Laser Designators, Harry Diamond Laboratories, HDL-TR-1860 (December 1982).

[2] Joseph V. Michalowicz, Analysis of the Laser-Guided Missile/Projectile versus Tank Counterfire Duel, Harry Diamond Laboratories, HDL-TR-1854 (May 1978).

III. PROBLEM VARIABLES. The essence of the encounter situation is best understood from the following parameters essential to the analysis.

Designator-tank range. For survivability reasons, the operator of the laser designator would prefer to operate at as great a distance from the tanks as possible, because the greater the range the less effective are the tank rounds counterfired against the GLD. However, there is an upper limit on this range imposed by the ability to hold the laser spot on the tank, particularly when the tank is moving. Also terrain conditions will often determine when the on-coming tank can first be designated. Consequently, a number of different designator-to-tank ranges are treated in this study.

Missile/projectile launch range. The flight time of the missile may affect the length of time the GLD must designate the target, if the target is illuminated for the entire missile flight. This is not the case for projectiles where laser designation is required only for the terminal part of the trajectory.

Tank response time. The time from laser alarm until the tank fires its first round, then a second round, then a third, etc, is one of the two critical time factors in the tank counterfire duel. The longer the tank response time, the more survivable the GLD. Time-to-first-fire and time-between-fires have been measured in various field tests, and probability distributions are fit to these data as the model is developed.

Designator-on time. The other critical time factor is the length of time that the target tank is illuminated by the GLD. The tradeoff between this time and the tank response time is the essence of the Counterfire Duel, for long designator-on times decrease the survivability of the GLD. The designator-on time may be reduced by special techniques such as offset designation, in which the GLD beams on a nearby object which reflects laser energy but does not trigger the tank's laser alarm, then switches to the tank in the final critical guidance phase of the weapon trajectory.

Single-shot kill probability (SSKP) for the laser-guided round against the tank. This parameter is of obvious importance and varies from one laser-guided weapon to another.

SSKP for tank round against laser designator. This parameter depends on the type of tank round fired, the tank-to-GLD range, the hardness of the FO position (e.g., in a foxhole, a bunker, or a forward observer vehicle (FOV)), and the ground slope at the GLD. The SSKP is also affected by the fire control and resulting aiming errors of the tank gun, which depend on whether the tank comes to a stop or fires on the move and whether the tank determines range visually or by means of a laser rangefinder. Permanent kill of the GLD is accomplished by either destroying the laser designator or disabling both the GLD operator and observer; temporary kill indicates either permanent kill or disabling only the operator.

Flight time of tank round. This parameter is determined by the type of round fired and the distance from the tank to the GLD. Since this time is nonzero, the possibility exists in the 1-on-1 duel that both the tank and the GLD will be killed.

Coordination of tank platoon. In the case of multiple missiles fired against a tank platoon, the degree to which the tanks in the platoon can alert their companion tanks when they are being designated may be a major factor in the outcome of the counterfire duel.

Time between missile fires. The time interval between successive laser-guided missile launches must be coordinated with the GLD operator and is dependent on the speed with which he can evaluate missile hit or miss and switch to the next target. It is expected that two or three missiles may be in flight simultaneously against a tank platoon.

To analyze the missile versus tank duel, the following exchange ratio is an appropriate performance measure:

$$ER = \frac{\text{Expected number of tanks killed}}{\text{Probability of GLD kill by tank counterfire}} . \qquad (1)$$

In the one missile versus one tank duel, this exchange ratio may be simplified to

$$ER = \frac{\text{Probability of tank kill}}{\text{Probability of GLD kill by tank counterfire}} . \qquad (2)$$

This performance measure provides a comparative description of the outcome of the missile versus tank encounter; large values of ER are favorable to the missile/GLD system, small values favorable to the tank. Note that a direct cost-effectiveness comparison would be difficult to formulate in this analysis, because the GLD bears a much greater significance than its actual unit cost since it is an essential part of an expensive weapon system.

IV. TANK RESPONSE TIME DISTRIBUTIONS. Various field tests have been conducted over the years to measure the rapidity with which a tank crew can recognize and fire upon a target which suddenly threatens the tank. Representative of the data chosen to portray these tank counterfire response times are the histograms in Figure 2, which depicts times for the tank to fire its first round (left-hand histogram) and times between fires (right-hand histogram).

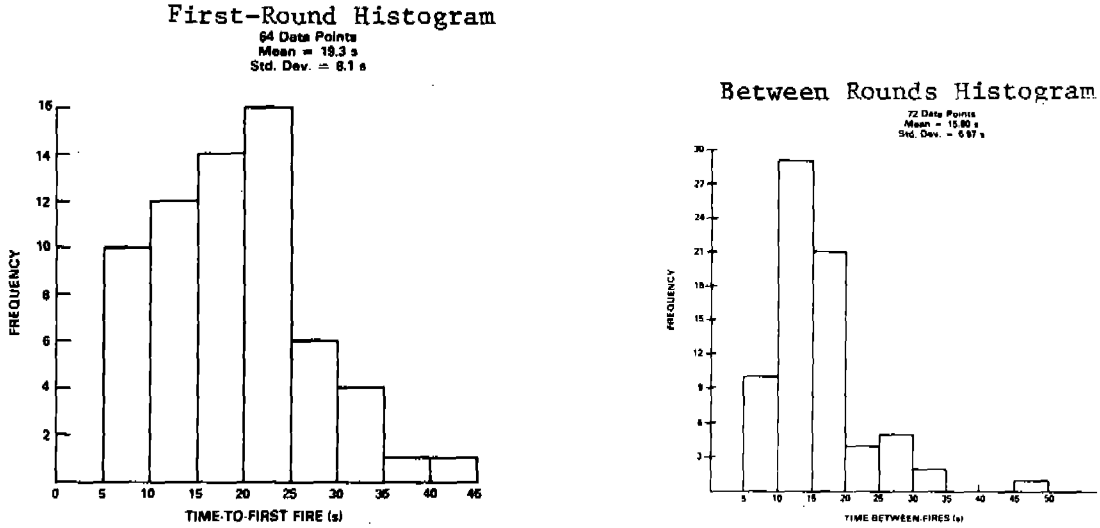


Figure 2. Typical Tank Response Time Data

In the past, log normal distributions were selected to fit such tank response time data. However, since the objective is to develop a tank counterfire model which is as realistic as possible, the tank must be permitted any number of counterfires during the designator-on period. That is, the number of rounds the tank can fire at the GLD will be limited only by the time in which the tank has to fire, and not by any arbitrary limit imposed to simplify the derivation of the relevant mathematical formulas. We shall see that this flexibility is not in consonance with the use of lognormal fits to the histograms.

We introduce the notation

t_n = time to nth tank fire,
Δt = time between tank fires.

Suppose a suitable probability density function $f(t_1)$ has been fit to the time-to-first-fire histogram and a density function $g(\Delta t)$ to the time-between-fires histogram. Under the assumption that the time-from-first-fire-to-second-fire distribution also represents that between any two successive fires, the time to the (k+1)st fire, for any $k \geq 1$, may be written

$$t_{k+1} = t_1 + \underbrace{\Delta t + \ . \ . \ . + \Delta t}_{\text{k independent choices of } \Delta t}$$

Since addition of random variables corresponds to convolution of their probability density functions, the density function for t_{k+1} is then given by the k-fold convolution

$$f(t_{k+1}) = f(t_1) * \underbrace{g(\Delta t) * \ . \ . \ . * g(\Delta t)}_{\text{k times}} \tag{3}$$

Each of these convolutions requires an integration, and so the problem appears to become computationally complicated beyond about 4 shots. The situation would be tractable if the distribution of Δt were reproductive (a distribution is defined to be reproductive if the sum of independent random variables each with such a distribution is a random variable which again has such a distribution). The normal distribution is a well-known reproductive distribution; however, this distribution does not have the proper shape to fit typical tank response-time histograms.

Finding a reproductive distribution to fit the time-between-fires histogram is crucial to the development of the present model; suitable candidate distributions are examined in table 1. The moment-generating functions are listed because the usual proof of reproductivity proceeds by showing that the moment-generating function for the sum of two independent random variables with the same type of distribution, which is the product of their two moment-generating functions, again corresponds to a distribution of that same type. In fact it turns out that the gamma, lognormal, and Rayleigh distributions all fit time-between-fires data reasonably well, but the

TABLE I. CANDIDATE TIME-BETWEEN-FIRES DISTRIBUTIONS

Density function	Formula	Fit	Moment-generating function	Reproductivity
Gamma	$f_{\lambda,\eta}(X) = \frac{\lambda^{\eta}}{\Gamma(\eta)} X^{\eta-1} e^{-\lambda X}$ For X>0	Good	$\left(1 - \frac{t}{\lambda}\right)^{-\eta}$	No, unless λ stays constant
Chi-square	$f_{\nu}(X) = \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)} X^{\frac{\nu}{2} - 1} e^{-\frac{X}{2}}$ For X>0	Special case of gamma	$(1 - 2t)^{-\nu/2}$	Yes
Lognormal	$f_{\mu,\sigma}(X) = \frac{1}{\sigma X\sqrt{2\pi}} e^{-(\ln X-\mu)^2/2\sigma^2}$ For X>0	Good	Doesn't exist	No
Rayleigh	$f_{\sigma}(X) = \left(\frac{X}{\sigma 2}\right)e^{-X^2/2\sigma^2}$ For $X \geq 0$	Good	$1 + te^{t^2\sigma^2/2}\left(\sigma\sqrt{\frac{\pi}{2}} + \int_{0}^{t\sigma^2} e^{-y^2/2\sigma^2}\,dy\right)$	No

lognormal and Rayleigh distributions are not reproductive. The gamma distribution is not in general reproductive, but this distribution is reproductive in the special case of repeated addition of a random variable to itself. And this is exactly the case here, where various values of Δt are summed.

Gamma distributions, fit to the tank response time data, then enable us to determine the density functions for t_{k+1}, for $k \geq 1$, with a single convolution.

With the notation

$$f(t_1) = f_{\lambda_1,\eta_1}(t_1)$$

$$g(\Delta t) = f_{\lambda,\eta}(\Delta t)$$

to denote the density functions of such gamma distributions, we can express the density functions for t_{k+1} as

$$f(t_{k+1}) = f(t_1) * \underbrace{g(\Delta t) * \ldots * g(\Delta t)}_{k \text{ times}}$$

$$= f_{\lambda_1,\eta_1}(t_1) * f_{\lambda,k\eta}(k\Delta t)$$

because of the partial reproductivity of the gamma distribution. Therefore, we have the formula

$$f(t_{k+1}) = \int_0^{t_{k+1}} f_{\lambda_1,\eta_1}(t_1) f_{\lambda,k\eta}(t_{k+1} - t_1) dt_1 \tag{4}$$

for each $k \geq 1$.

The parameters λ and η of the gamma distribution fits to the response time histograms are obtained by equating η/λ to the test data mean and η/λ^2 to the test data variance. In this way the following gamma distribution fits are derived:

(1) time-to-first-fire

$$\lambda_1 = 0.269$$

$$\eta_1 = 5.262$$

$$f(t_1) = (2.783 \times 10^{-5})\, t_1^{4.262} \exp(-0.269 t_1)$$

This fit passes the chi-square test, since a value of

$$\chi^2 = 4.33$$

is calculated with 3 degrees of freedom, which compares favorably with the critical value at the 5-percent level (type I error)

$$\chi_o^2 = 7.815$$

(2) time-between-fires

$$\lambda = 0.3273$$

$$\eta = 5.204$$

$$g(\Delta t) = (9.123 \times 10^{-5})\Delta t^{4.204}\exp(-0.3273\Delta t)$$

This fit also passes the chi-square test, at least at the 2-percent level.

These gamma distribution fits are shown in figures 3 and 4.

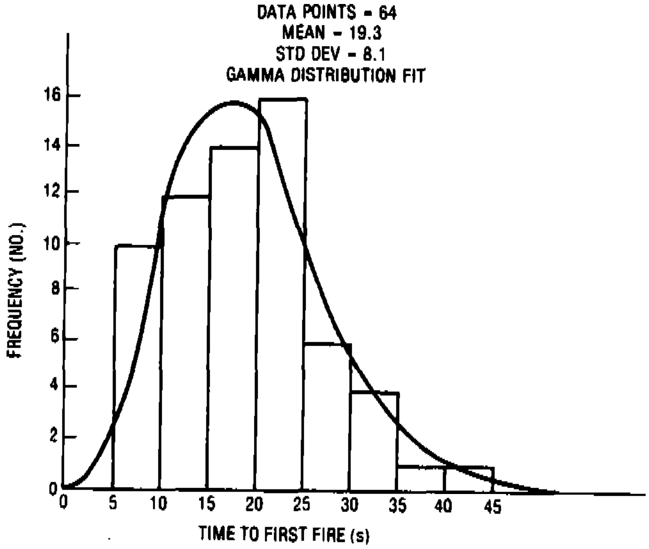


Figure 3. Gamma distribution fit to first-round histogram.

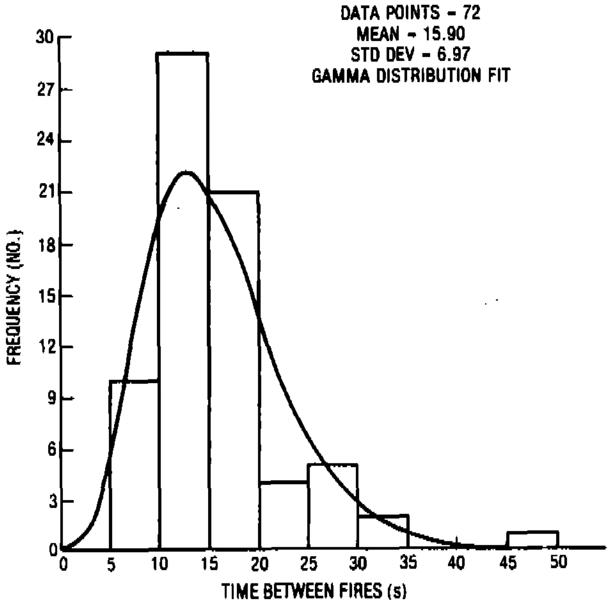


Figure 4. Gamma distribution fit to between-rounds histogram.

Substitution of the gamma distribution parameters into equation (4) then gives the formulas for the time-to-n^th-fire probability densities:

$$f(t_1) = (2.783 \times 10^{-5})t_1^{4.262}\exp(-0.269t_1) \qquad (5)$$

$$f(t_{k+1}) = (2.783 \times 10^{-5})(0.00299)^k\exp(-0.3273t_{k+1}) \cdot$$

$$\int_0^{t_{k+1}} t_1^{4.262}\frac{(t_{k+1} - t_1)^{5.204k-1}}{\Gamma(5.204k)} \exp(0.0583t_1)\, dt_1 \qquad (6)$$

for $k \geq 1$.

Graphs of some of these time-to-n^th-fire probability density functions are presented in figures 5 through 9.

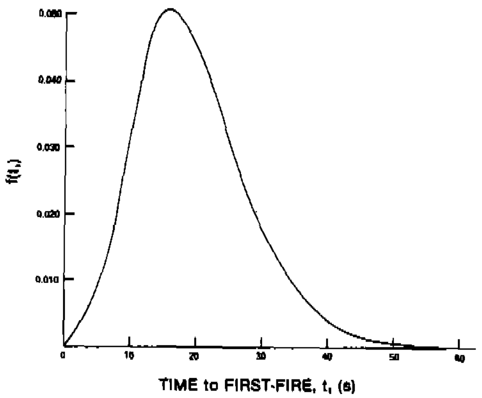


Figure 5. Probability density for time-to-first-fire

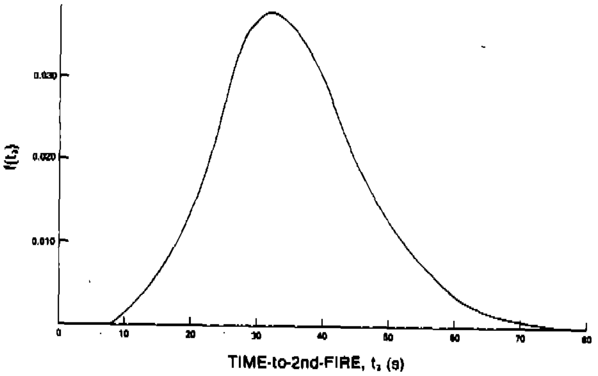


Figure 6. Probability density for time-to-second-fire

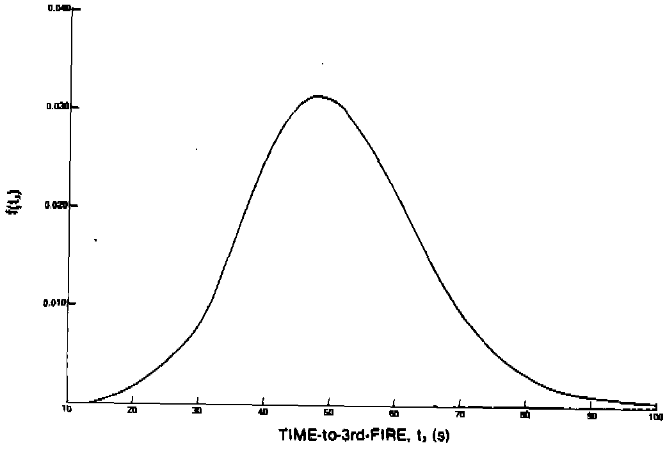


Figure 7. Probability density for time-to-third-fire

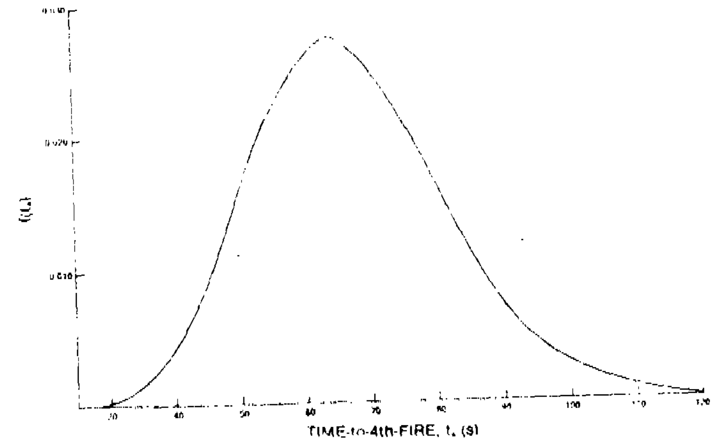


Figure 8. Probability density for time-to-fourth-fire

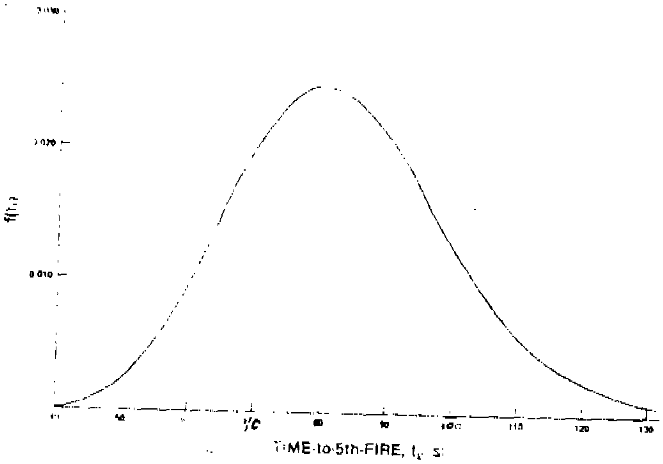


Figure 9. Probability density for time-to-fifth-fire

V. TANK COUNTERFIRE MODELS. With the tank response time distributions derived in the previous section, analytic models can now be developed for the 1-on-1 duel and the 3-on-3 encounter. Formulas will be given expressing the probability of GLD destruction and the expected number of tanks killed in each encounter. Any number of tank counterfires may occur during the designator-on time and the flight time of the tank round is explicitly considered. Coordination of counterfire from the tank platoon is also treated.

Several ground rules are established in developing the model, although different assumptions could be readily incorporated. Line-of-sight between the GLD and the tank is maintained throughout the duel encounter; this is not unrealistic, especially when the tank or tanks stop to fire. If the tank laser alarm system is operating properly, it is assumed to detect the existence of the laser spot as soon as the tank is illuminated (if there were a known lag time, it would be subtracted from the designator-on time). Once the laser alarm sounds, the tank crew is assumed to devote full attention to defeating the GLD rather than pursuing its original mission. If the tank destroys the GLD, the laser-guided round is rendered harmless to the tank. Suppression is not played, so tank rounds which miss the GLD do not disturb the GLD operator's illumination of the tank target and hence do not disrupt the operation of the laser-guided weapon system. Tanks may fire either at a stop or on the move, and they may adjust their aim between rounds; these possibilities can be handled by using the appropriate SSKP data. The assumption that the tanks fire at the GLD only during the designator-on time serves as the end-of-game criterion.

One-on-One Duel. The duel between a single laser-guided missile directed against an individual tank by a GLD, with that tank counterfiring against the GLD, will be treated first. The following notation will be needed to express the resulting probability formulas.

T_1 = designator-on time = time from laser alarm to missile hit.
T_2 = flight time of tank-fired round
P_D = probability that the tank detects the laser designation and fires at the GLD
P_{TK} = SSKP by laser-guided missile
P_{GK} = SSKP by tank round
P_n = probability that the tank fires the nth round in T_1 seconds
= Prob $(t_n \leq T_1)$
P_n^* = probability that the tank fires the nth round in $T_1 - T_2$ seconds
= Prob $(t_n \leq T_1 - T_2)$

Since the tank round is unguided after fire, both the tank and the GLD could be destroyed if the tank fires a round at the GLD less than T_2 seconds before missile hit. It is for this reason that the distinction between P_n and P_n^* is drawn.

The formulas which govern the 1-on-1 duel may be written as follows:

$$P = \text{Prob (tank killed by missile)} \tag{7}$$

$$= P_{TK}\left[1 - P_D P_1^* + P_D \sum_{n=1}^{\infty} (P_n^* - P_{n+1}^*)(1 - P_{GK})^n\right]$$

$$Q = \text{Prob (GLD killed by tank)} \tag{8}$$

$$= P_D P_{GK} \sum_{n=1}^{\infty} P_n (1 - P_{GK})^{n-1}$$

$$ER = \frac{P}{Q} \tag{9}$$

Formula (7) for tank kill is derived by expressing the event that the tank is killed by the laser-guided missile as a sum of disjoint events, where the nth term in the summation corresponds to the event in which the tank has fired exactly n rounds before being destroyed by the missile but all miss. Likewise, formula (8) for GLD kill is constructed from a sum of disjoint events, with the nth term in the summation representing the event in which the tank fires n rounds before missile hit and the nth round is the one that destroys the GLD. The exchange ratio ER is then given by the quotient of these two probabilities.

Three Missiles versus Tank Platoon. Derivation of the probability formulas is considerably more complex for the counterfire scenario of three laser-guided missiles rapid-fired under the control of one GLD against a platoon of three tanks. In this case the GLD switches to another tank upon tank kill and two or three missiles may be in the air at one time in an attack sequence. An important factor in the analysis of this scenario is the degree to which the tanks can alert their companion tanks to the presence of the laser designator. Two cases will be considered: an uncoordinated platoon, and a perfectly coordinated platoon. The following notation, which extends that used for the one-on-one duel, will be needed in developing the formulas (T denotes a time variable):

$P_n(T)$ = probability of nth tank fire in T seconds

$P_n^*(T)$ = probability of nth tank fire in T - T_2 seconds

$P^*(T)$ = probability that all tank rounds fired in T - T_2 seconds miss

$$= 1 - P_D P_1^*(T) + P_D \sum_{n=1}^{\infty} \left(P_n^*(T) - P_{n+1}^*(T)\right)\left(1 - P_{GK}\right)^n$$

$P(T)$ = probability that all tank rounds fired in T seconds miss

$$= 1 - P_D P_1(T) + P_D \sum_{n=1}^{\infty} \left(P_n(T) - P_{n+1}(T)\right)\left(1 - P_{GK}\right)^n$$

$Q(T)$ = probability that the tank fires a GLD-killing round in T seconds

$$= P_D P_{GK} \sum_{n=1}^{\infty} P_n(T)\left(1 - P_{GK}\right)^{n-1}$$

$Q^*(T)$ = probability that the tank fires a GLD-killing round in T - T_2 seconds

$$= P_D P_{GK} \sum_{n=1}^{\infty} P_n^*(T)\left(1 - P_{GK}\right)^{n-1}$$

T_3 = length of time that GLD illuminates companion tanks

T_4 = time between missile arrivals

T_5 = companion tank reaction time delay.

$P^{(i)}$ = probability that i tanks are killed

Q = probability that GLD is killed

$Q^{(i)}$ = probability that GLD is killed and i tanks are killed

It should be clear that

$P^{(i)} - Q^{(i)}$ = probability that GLD survives and i tanks are killed,

and that if there are m target tanks (e.g., m = 3 for a tank platoon), then

$$\sum_{i=0}^{m} P^{(i)} = 1 \quad ,$$

$$\sum_{i=0}^{m} Q^{(i)} = Q \quad ,$$

$$E = \text{expected number of tanks killed} = \sum_{i=0}^{m} iP^{(i)} \quad .$$

T_1 will now be interpreted as the time from laser alarm of the first tank designated to the arrival of the first missile, to be consistent with the usage of the one-on-one duel. It is easily shown that for any value of T

$$P(T) + Q(T) = 1$$

$$P*(T) + Q*(T) = 1 \ .$$

The convention is adopted that $P^*(T) = 1$ and $Q^*(T) = 0$ when $T < T_2$.

Uncoordinated Tank Platoon. First, suppose the designated tanks do not, or are unable to, alert companion tanks in the platoon to the presence of the GLD. In this case, the tanks counterfire only upon being designated themselves, so the logical GLD strategy is to beam on a tank until it is killed, then switch to the next tank and designate it for T_3 seconds before the impact of the next missile. The time, T_4, between missile fires has to allow for this switching time. If a missile misses its intended target, the GLD still maintains its designation of the target tank instead of switching to another tank at this point, because switching would incur counterfire from another tank in addition to the continued counterfire from the tank originally designated. The game ends when the last missile reaches its target; the GLD switches off and no further tank rounds are fired (however, the effect of a tank round already in flight is included).

The equations governing this situation are presented in figure 10. The assumption is made that the time between missile arrivals is greater than the flight time of the tank round; that is

$$T_4 \geq T_2$$

This assumption is valid in most tank counterfire cases. Modifications have to be made to the formulas if this inequality is reversed; such formulas are included in another report[3].

$$
\begin{aligned}
P^{(3)} &= \text{Prob (3 tanks killed)} \\
&= P(T_1)P(T_3)P^*(T_3)P_{TK}^3
\end{aligned}
$$

$$
\begin{aligned}
P^{(2)} &= \text{Prob (exactly 2 tanks killed)} \\
&= P(T_1)\Big(\big[Q(T_3) - Q^*(T_3)\big] + P(T_3)\big[1 - P^*(T_3)P_{TK}\big]\Big)P_{TK}^2 \\
&\quad + P(T_1)P^*(T_3 + T_4)(1 - P_{TK})P_{TK}^2 \\
&\quad + P(T_1 + T_4)P^*(T_3)(1 - P_{TK})P_{TK}^2
\end{aligned}
$$

$$
\begin{aligned}
P^{(1)} &= \text{Prob (exactly one tank killed)} \\
&= \big[Q(T_1) - Q^*(T_1)\big]P_{TK} \\
&\quad + P(T_1)\Big(Q^*(T_3) + \big[Q^*(T_3 + T_4) - Q^*(T_3)\big](1 - P_{TK}) + P^*(T_3 + T_4)(1 - P_{TK})^2\Big)P_{TK} \\
&\quad + \big[Q(T_1 + T_4) - Q^*(T_1 + T_4)\big](1 - P_{TK})P_{TK} \\
&\quad + P(T_1 + T_4)\big[1 - P^*(T_3)P_{TK}\big](1 - P_{TK})P_{TK} \\
&\quad + P^*(T_1 + 2T_4)(1 - P_{TK})^2P_{TK}
\end{aligned}
$$

Figure 10. Formulas: Three missiles versus uncoordinated tank platoon.

[3] Joseph V. Michalowicz, Analysis of the Laser-Guided Maverick versus Tank Counterfire Duel, Harry Diamond Laboratories, HDL-TR-1909 (June 1980).

$$\begin{aligned} P^{(0)} &= \text{Prob (no tanks killed)} \\ &= Q^*(T_1) \\ &+ \left[Q^*(T_1 + T_4) - Q^*(T_1)\right](1 - P_{TK}) \\ &+ \left[Q^*(T_1 + 2T_4) - Q^*(T_1 + T_4)\right](1 - P_{TK})^2 \\ &+ P^*(T_1 + 2T_4)(1 - P_{TK})^3 \end{aligned}$$

$$\begin{aligned} Q &= \text{Prob (GLD killed)} \\ &= P(T_1)P(T_3)Q(T_3)P_{TK}^2 \\ &+ P(T_1)\Big(Q(T_3) + \left[Q(T_3 + T_4) - Q(T_3)\right](1 - P_{TK})\Big)P_{TK} \\ &+ P(T_1 + T_4)Q(T_3)(1 - P_{TK})P_{TK} \\ &+ Q(T_1) + \left[Q(T_1 + T_4) - Q(T_1)\right](1 - P_{TK}) + \left[Q(T_1 + 2T_4) - Q(T_1 + T_4)\right](1 - P_{TK})^2 \end{aligned}$$

$$Q^{(3)} = P(T_1)P_{TK}P(T_3)P_{TK}[Q(T_3) - Q^*(T_3)]P_{TK}$$

$$\begin{aligned} Q^{(2)} &= P(T_1)P_{TK}P(T_3)P_{TK}\Big(Q^*(T_3) + (1 - P_{TK})[Q(T_3) - Q^*(T_3)]\Big) + P(T_1)P_{TK}P_{TK}[Q(T_3) - Q^*(T_3)] \\ &+ P(T_1)P_{TK}(1 - P_{TK})P_{TK}[Q(T_3 + T_4) - Q^*(T_3 + T_4)] \\ &+ (1 - P_{TK})P(T_1 + T_4)P_{TK}P_{TK}[Q(T_3) - Q^*(T_3)] \end{aligned}$$

$$\begin{aligned} Q^{(1)} &= P_{TK}\Big([Q(T_1) - Q^*(T_1)] + P(T_1)Q^*(T_3) + P(T_1)(1 - P_{TK})[Q^*(T_3 + T_4) - Q^*(T_3)] \\ &+ P(T_1)(1 - P_{TK})^2[Q(T_3 + T_4) - Q^*(T_3 + T_4)]\Big) \\ &+ (1 - P_{TK})P_{TK}\Big([Q(T_1 + T_4) - Q^*(T_1 + T_4)] + P(T_1 + T_4)Q^*(T_3) \\ &+ P(T_1 + T_4)(1 - P_{TK})[Q(T_3) - Q^*(T_3)]\Big) \\ &+ (1 - P_{TK})^2P_{TK}[Q(T_1 + 2T_4) - Q^*(T_1 + 2T_4)] \end{aligned}$$

$$\begin{aligned} Q^{(0)} &= Q^*(T_1) + (1 - P_{TK})[Q^*(T_1 + T_4) - Q^*(T_1)] + (1 - P_{TK})^2[Q^*(T_1 + 2T_4) - Q^*(T_1 + T_4)] \\ &+ (1 - P_{TK})^3[Q(T_1 + 2T_4) - Q^*(T_1 + 2T_4)] \end{aligned}$$

$$\begin{aligned} E &= \text{Expected number of tanks killed} \\ &= 3P^{(3)} + 2P^{(2)} + P^{(1)} \end{aligned}$$

$$\text{Exchange ratio} = \frac{E}{Q}$$

Figure 10. Formulas: Three missiles versus uncoordinated tank platoon (cont'd).

Each of the formulas in Figure 10 is derived from a a sum of disjoint events as for the 1-on-1 duel, but it is clear that the complexity of the formulas has greatly increased in the multiple missile/multiple tanks case. We choose one example, say the formula for $P^{(1)}$, to show how these equations are derived. Now $P^{(1)}$ is the probability that exactly one tank is killed; the disjoint events and their corresponding probabilities, which add up to give the formula for $P^{(1)}$, are shown in Figure 11.

Event	Probability
1st missile kills 1st tank but an in-flight round from the 1st tank kills the GLD.	$P_{TK}[Q(T_1)-Q^*(T_1)]$
1st missile kills 1st tank and either the GLD is killed by 2nd tank before 2nd missile arrives or before the 3rd missile if the 2nd missile misses, or both the 2nd and 3rd missiles miss.	$P(T_1)P_{TK}(Q^*(T_3) + (1-P_{TK})[Q^*(T_3+T_4)-Q^*(T_3)] + P^*(T_3+T_4)\ (1-P_{TK})^2)$
1st missile misses but 2nd missile kills 1st tank and an in-flight round from the 1st tank kills the GLD.	$(1-P_{TK})P_{TK}[Q(T_1+T_4) - Q^*(T_1+T_4)]$
1st missile misses but 2nd missile kills 1st tank and either 2nd tank counterfire kills the GLD or the 3rd missile misses.	$P(T_1+T_4)(1-P_{TK})P_{TK}\ [Q^*(T_3) + P^*(T_3)\ (1-P_{TK})]$
1st and 2nd missiles miss but the 3rd missile kills the 1st tank	$P^*(T_1+2T_4)(1-P_{TK})^2P_{TK}$

Figure 11. Disjoint events used to derive formula for $P^{(1)}$

Coordinated Tank Platoon. On the other hand, suppose that the tank platoon is coordinated in the sense that upon laser alarm the designated tank alerts its companion tanks and they begin counterfire as well, after a time delay, T_5, required either to receive the information from the designated tank or to observe its reaction to recognition of the laser illumination (e.g., stopping, slewing of the turret). Since in practice $T_1 > T_5$, all tanks will have begun counterfiring at the GLD before the arrival of the first missile. The GLD is assumed to use the same designation strategy as in the case of the uncoordinated tank platoon, since it is of no benefit to switch targets until the kill is observed. The accuracy, and hence the SSKP, of the companion tank rounds will be assumed to be the same as the designated tank.

In figure 12 the equations are presented for this case; derivations are again based on sums of disjoint events. The assumption that $T_4 \geq T_2$ remains in effect for these formulas:

$$\begin{aligned}
P^{(3)} &= \text{Prob (3 tanks killed)} \\
&= P(T_1)P(T_1 + T_4 - T_5)P^*(T_1 + 2T_4 - T_5)P_{TK}^3
\end{aligned}$$

$$\begin{aligned}
P^{(2)} &= \text{Prob (exactly 2 tanks killed)} \\
&= P(T_1)\left[Q(T_1 + T_4 - T_5) - Q^*(T_1 + T_4 - T_5)\right]P^*(T_1 + T_4 - T_5)P_{TK}^2 \\
&+ P(T_1)P(T_1 + T_4 - T_5)\left[Q^*(T_1 + 2T_4 - T_5) - Q^*(T_1 + T_4 - T_5)\right]P_{TK}^2 \\
&+ P(T_1)P(T_1 + T_4 - T_5)P^*(T_1 + 2T_4 - T_5)P_{TK}^2(1 - P_{TK}) \\
&+ P(T_1)\left[P^*(T_1 + 2T_4 - T_5)\right]^2 P_{TK}^2(1 - P_{TK}) \\
&+ P(T_1 + T_4)\left[P^*(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})P_{TK}^2
\end{aligned}$$

$$\begin{aligned}
P^{(1)} &= \text{Prob (exactly one tank killed)} \\
&= \left[Q(T_1) - Q^*(T_1)\right]\left[P^*(T_1 - T_5)\right]^2 P_{TK} \\
&+ P(T_1)\left(\left[P^*(T_1 - T_5)\right]^2 - \left[P^*(T_1 + T_4 - T_5)\right]^2\right)P_{TK} \\
&+ P(T_1)\left(\left[P^*(T_1 + T_4 - T_5)\right]^2 - \left[P^*(T_1 + 2T_4 - T_5)\right]^2\right)P_{TK}(1 - P_{TK}) \\
&+ P(T_1)\left[P^*(T_1 + 2T_4 - T_5)\right]^2 P_{TK}(1 - P_{TK})^2 \\
&+ \left[Q(T_1 + T_4) - Q^*(T_1 + T_4)\right]\left[P^*(T_1 + T_4 - T_5)\right]^2(1 - P_{TK})P_{TK} \\
&+ P(T_1 + T_4)\left(\left[P^*(T_1 + T_4 - T_5)\right]^2 - \left[P^*(T_1 + 2T_4 - T_5)\right]^2\right)(1 - P_{TK})P_{TK} \\
&+ P(T_1 + T_4)\left[P^*(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})^2 P_{TK} \\
&+ P^*(T_1 + 2T_4)\left[P^*(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})^2 P_{TK}
\end{aligned}$$

Figure 12. Formulas: three missiles versus coordinated tank platoon.

$$P^{(0)} = \text{Prob (no tanks killed)}$$

$$= 1 - P*(T_1)\left[P*(T_1 - T_5)\right]^2$$

$$+ \left(P*(T_1)\left[P*(T_1 - T_5)\right]^2 - P*(T_1 + T_4)\left[P*(T_1 + T_4 - T_5)\right]^2\right)(1 - P_{TK})$$

$$+ \left(P*(T_1 + T_4)\left[P*(T_1 + T_4 - T_5)\right]^2 - P*(T_1 + 2T_4)\left[P*(T_1 + 2T_4 - T_5)\right]^2\right)(1 - P_{TK})^2$$

$$+ P*(T_1 + 2T_4)\left[P*(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})^3$$

$$Q = \text{Prob (GLD killed)}$$

$$= 1 - \text{Prob (GLD not killed)}$$

$$= 1 - \Big(P(T_1)P(T_1 + T_4 - T_5)P(T_1 + 2T_4 - T_5)P_{TK}^2$$

$$+ P(T_1)\left[P(T_1 + 2T_4 - T_5)\right]^2P_{TK}(1 - P_{TK})$$

$$+ P(T_1 + T_4)\left[P(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})P_{TK}$$

$$+ P(T_1 + 2T_4)\left[P(T_1 + 2T_4 - T_5)\right]^2(1 - P_{TK})^2\Big)$$

$$Q^{(3)} = P(T_1)P_{TK}P(T_1 + T_4 - T_5)P_{TK}[Q(T_1 + 2T_4 - T_5) - Q^*(T_1 + 2T_4 - T_5)]P_{TK}$$

$$Q^{(2)} = P(T_1)P_{TK}P_{TK}\ \Big([Q(T_1 + T_4 - T_5) - Q^*(T_1 + T_4 - T_5)]P^*(T_1 + T_4 - T_5)$$

$$+ P(T_1 + T_4 - T_5)[Q^*(T_1 + 2T_4 - T_5) - Q^*(T_1 + T_4 - T_5)]$$

$$+ P(T_1 + T_4 - T_5)(1 - P_{TK})[Q(T_1 + 2T_4 - T_5) - Q^*(T_1 + 2T_4 - T_5)]\Big)$$

$$+ P(T_1)P_{TK}(1 - P_{TK})P_{TK}\ \Big([P^*(T_1 + 2T_4 - T_5)]^2 - [P(T_1 + 2T_4 - T_5)]^2\Big)$$

$$+ (1 - P_{TK})P(T_1 + T_4)P_{TK}P_{TK}\ \Big([P^*(T_1 + 2T_4 - T_5)]^2 - [P(T_1 + 2T_4 - T_5)]^2\Big)$$

$$Q^{(1)} = P_{TK}\ \Big[[Q(T_1) - Q^*(T_1)][P^*(T_1 - T_5)]^2 + P(T_1)\ \Big([P^*(T_1 - T_5)]^2 - [P^*(T_1 + T_4 - T_5)]^2\Big)$$

$$+ P(T_1)(1 - P_{TK})\ \Big([P^*(T_1 + T_4 - T_5)]^2 - [P^*(T_1 + 2T_4 - T_5)]^2\Big)$$

$$+ P(T_1)(1 - P_{TK})^2\ \Big([P^*(T_1 + 2T_4 - T_5)]^2 - [P(T_1 + 2T_4 - T_5)]^2\Big)\Big]$$

$$+ (1 - P_{TK})P_{TK}\ \Big[[Q(T_1 + T_4) - Q^*(T_1 + T_4)][P^*(T_1 + T_4 - T_5)]^2$$

$$+ P(T_1 + T_4)\ \Big([P^*(T_1 + T_4 - T_5)]^2 - [P^*(T_1 + 2T_4 - T_5)]^2\Big)$$

$$+ P(T_1 + T_4)(1 - P_{TK})\ \Big([P^*(T_1 + 2T_4 - T_5)]^2 - [P(T_1 + 2T_4 - T_5)]^2\Big)\Big]$$

$$+ (1 - P_{TK})^2P_{TK}\ \Big(P^*(T_1 + 2T_4)[P^*(T_1 + 2T_4 - T_5)]^2 - P(T_1 + 2T_4)[P(T_1 + 2T_4 - T_5)]^2\Big)$$

$$Q^{(0)} = 1 - P^*(T_1)[P^*(T_1 - T_5)]^2 + (1 - P_{TK})\ \Big(P^*(T_1)[P^*(T_1 - T_5)]^2 - P^*(T_1 + T_4)[P^*(T_1 + T_4 - T_5)]^2\Big)$$

$$+ (1 - P_{TK})^2\ \Big(P^*(T_1 + T_4)[P^*(T_1 + T_4 - T_5)]^2 - P^*(T_1 + 2T_4)[P^*(T_1 + 2T_4 - T_5)]^2\Big)$$

$$+ (1 - P_{TK})^3\ \Big(P^*(T_1 + 2T_4)[P^*(T_1 + 2T_4 - T_5)]^2 - P(T_1 + 2T_4)[P(T_1 + 2T_4 - T_5)]^2\Big)$$

$$E = \text{Expected number of tanks killed}$$

$$= 3P^{(3)} + 2P^{(2)} + P^{(1)}$$

$$\text{Exchange ratio} = \frac{E}{Q}$$

Figure 12. Formulas: three missiles versus coordinated tank platoon. (cont'd)

The model has been developed thus far under the assumption that the tank round SSKP, P_{GK}, remains constant for each fire. However, the tanker may be able to adjust his aim based on the result of his first fire and thus improve the SSKP for succeeding fires, or he may load a different round for subsequent fires from the one he was originally carrying in his main gun, or he may estimate the range to the GLD either visually or with a laser rangefinder and thus increase the accuracy and lethality of the rounds fired at the expense of a small time delay for ranging. All of these sophistications can be added to the above formulas with little difficulty.

VI. SAMPLE RESULTS AND CONCLUSIONS. The tank counterfire duel models which have been presented were exercised for many combinations of parameters in order to provide the encounter outcomes for the SSTF force-on-force analysis. One way to graphically display the results of a counterfire duel is demonstrated in Figure 13 which is based upon a sample set of input parameters. The graph shows the expected number of tanks killed (solid curves) and the probability of GLD kill (dashed curves) in a 3-on-3 encounter over a range of tank-to-GLD distances, with the GLD located in either a foxhole, a Forward Observer Vehicle or a bunker.

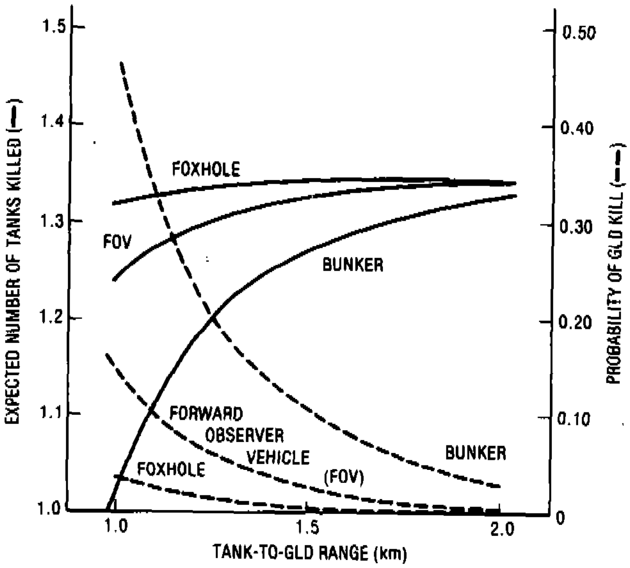


Figure 13. Sample Counterfire Duel Results

These same results are represented in tabular form, for a particular intermediate tank-to-GLD range, in Table 2.

Table 2. Sample Tabular Counterfire Duel Results

Tank-to-GLD range (km)	GLD Position	$P^{(0)}$	$P^{(1)}$	$P^{(2)}$	$P^{(3)}$	E	$Q^{(0)}$	$Q^{(1)}$	$Q^{(2)}$	$Q^{(3)}$	Q	ER
1.5	Foxhole	0.170	0.409	0.331	0.090	1.341	0.002	0.002	0.000	0.000	0.004	335
1.5	Bunker	0.203	0.405	0.308	0.084	1.273	0.058	0.043	0.007	0.000	0.108	12
1.5	FOV	0.176	0.409	0.327	0.088	1.327	0.014	0.010	0.002	0.000	0.026	51

The exchange ratios obtained in this sample calculation indicate that the foxhole position is the most survivable and effective GLD position, while the bunker is the least survivable to tank counterfire. No actual conclusions should be drawn from this sample calculation since these results are highly dependent on the choice of input parameters, such as the type of tank rounds fired, but this same line of reasoning would be used with actual data to draw survivability conclusions.

Incorporating the counterfire duel methodology into the SSTF war game scenario provided the data needed to determine whether or not the GLD is survivable on the battlefield. Due to classification, the conclusion of the SSTF analysis cannot be presented in this paper. However, it should be mentioned that the counterfire duel methodology was also very useful in testing the sensitivity of the SSTF conclusions to variations in the many input parameters required in the study -- such as type of tank rounds fired, motion of tank, handoff time between tanks, designation mode, slope at GLD position, and laser-guided weapon rate of fire -- to determine which were critical. Such sensitivity analyses are often instrumental in the development of weapon system improvements.

It is anticipated that the analytic model presented in this paper is sufficiently general to have applications to many types of duel encounters where response time is the critical factor. We hope that the reader will find the model useful in such situations; further details can be obtained from the author.

Algorithm for Calculating Unit Separation Distances

Timothy M. Geipe
Joseph V. Michalowicz

Harry Diamond Laboratories, USA ERADCOM

Abstract

Battlefield units adjacent to a targeted unit must maintain some separation distance to avoid collateral damage. A tabular algorithm for determining such separation distances based on unit damage criteria and weapon delivery errors for several confidence levels is presented. The algorithm depends on a numerical technique for integrating a two-dimensional weapon burst distribution function over some base region and an iterative technique to obtain separation distances given other known parametes. The use of these numerical techniques is discussed along with several current applications of the algorithm.

1. INTRODUCTION

On the tactical nuclear battlefield, if the enemy (Red) is able to accurately determine the location of a high-priority, friendly (Blue) unit, he is expected to fire a nuclear round of sufficient yield and accuracy to destroy the target with a high degree of confidence. That unit is effectively lost, so the important question concerning Blue survivability is the "bonus damage" produced by that nuclear fire on adjacent, nontargeted units.

A typical problem is the calculation of the desired minimum separation distance between "neighbor" units. For a given confidence level, C, this distance is determined as that at which some specified environment or environments, due to a nuclear burst at the target unit, is exceeded at the adjacent neighbor unit with a probability of only 1 - C. This report presents a handy tabular algorithm to calculate these specific environmental criteria for various confidence levels.

Applications of the methodology are developed for several examples: (1) command post survivability, (2) weapon employment, and (3) trade-off between hardening and operational deployment.

2. METHODOLOGY

Let us suppose that the target unit is located at ground zero (GZ) and that the burst point (more precisely, the projection on the ground of the point at which detonation occurs) has a two-dimensional normal distribution about GZ with density function

$$p(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(x^2+y^2\right)/2\sigma^2} , \qquad (1)$$

where the standard deviation σ (of the marginal distributions) results from both weapon-delivery error and target-location error. Assume that the adjacent unit is located at a separation distance S from GZ as shown in figure 1. For a given confidence level, C (e.g., C = 0.90), the distance D is calculated at which the probability is only 1 - C that the burst point of the round falls within D of the adjacent unit, from the following formula:

$$2\int_0^D \int_{S-\sqrt{D^2-y^2}}^{S+\sqrt{D^2-y^2}} p(x,y)\,dx\,dy = 1 - C \quad . \qquad (2)$$

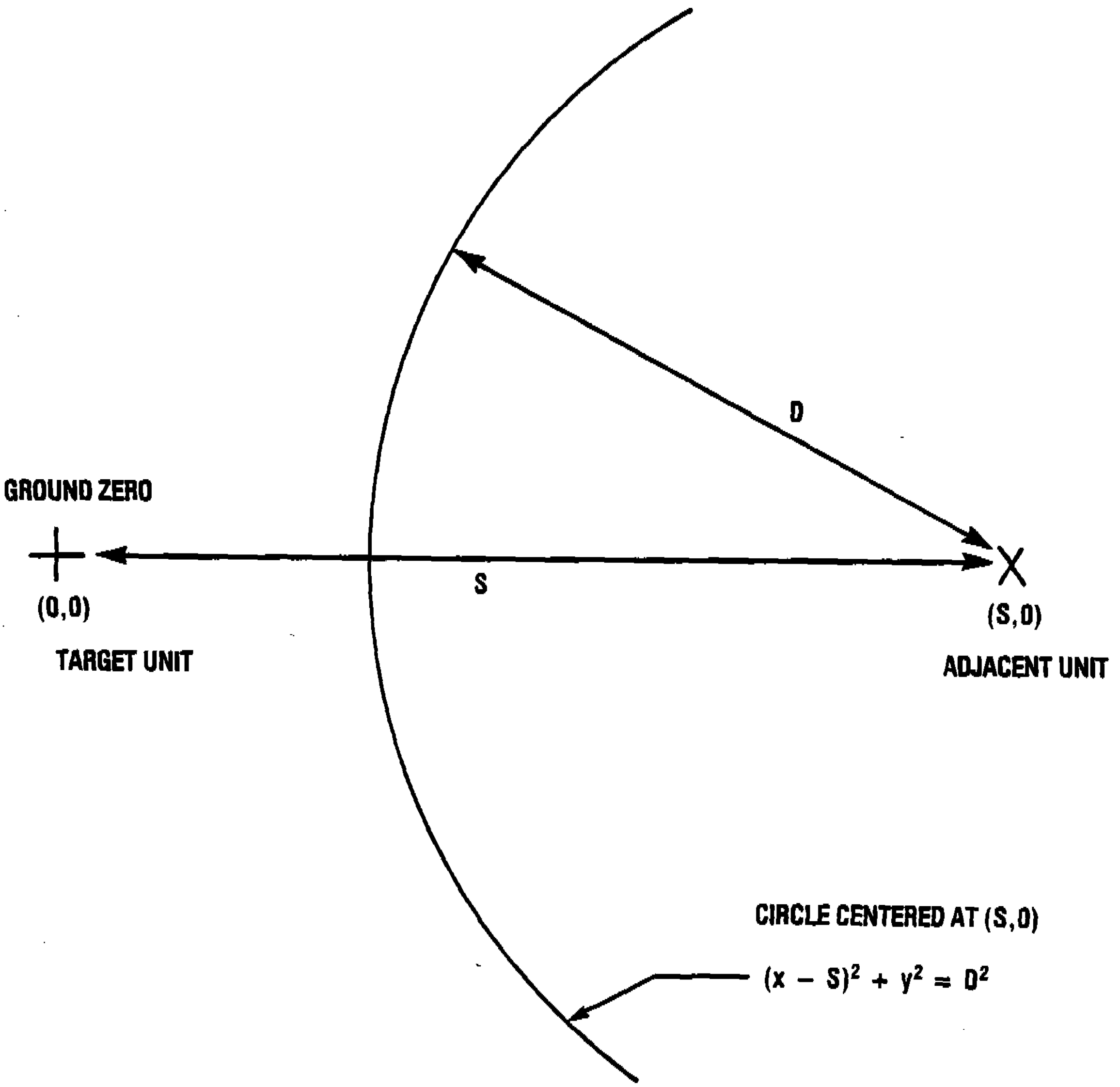


Figure 1. Geometry of Units.

Solving equation (2) for D as a function of C clearly involves a process of both integration and iteration. A "stack of disks" method is used to perform the double integration of the integrand, whose graph yields the surface shown in figure 2.

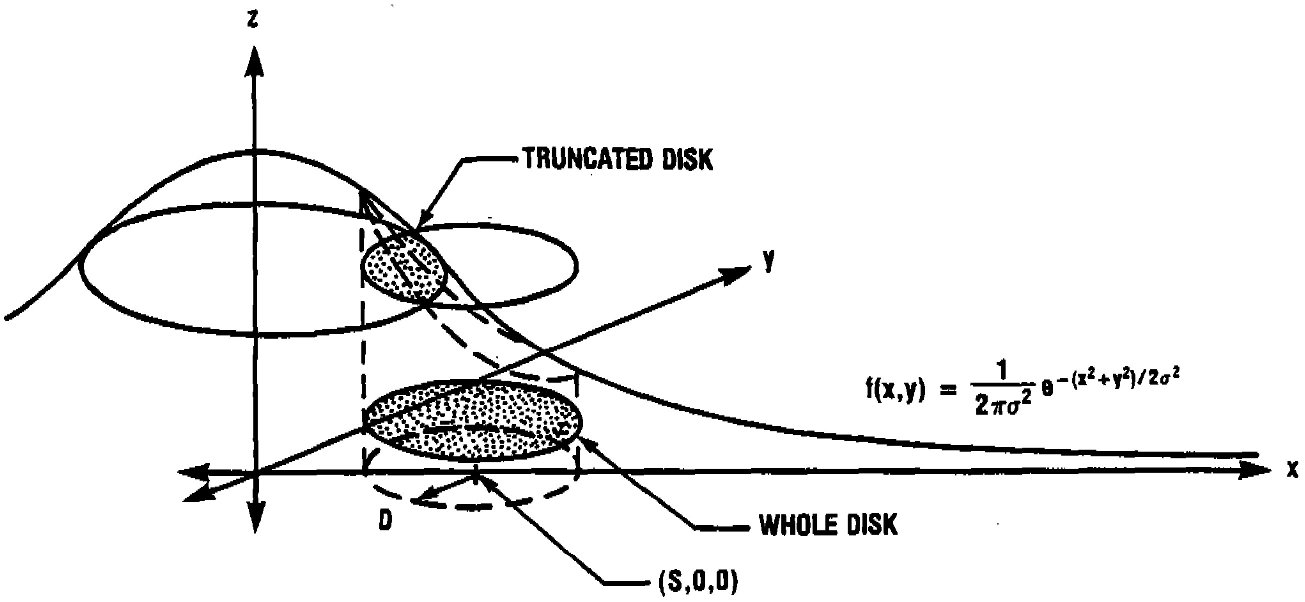


Figure 2. Integration Surface.

The double integral is represented by the volume of the figure formed by the intersection of the surface with a cylinder of radius D, centered at the point (S, O, O). This volume is calculated by slicing the figure into disks, each of which has a cross section which is either a circle or the intersection of two circles, and adding up the volume of all of the disks. The iterative process required to solve for D for a given C is a "telescoping" technique akin to that used for finding roots of polynomial equations.

Fomulas have been derived[1] for calculating the following nuclear environments as a function of range:

Initial Nuclear Radiation

- Total dose
- Transient Radiation Effects on Electronics
  - Neutron Fluence
  - Peak gamma-dose rate

Thermal Radiation

- Radiant Energy
- Maximum Thermal Irradiance

[1] W. E. Sweeney, Jr., Cyrus G. Moazed, and John S. Wicklund "Nuclear Weapons Environments for Vulnerability Assessments to Support Tactical Nuclear Warfare Studies (U)," Harry Diamond Laboratories TM-77-4 (June 1977). (CONFIDENTIAL)

Blast

- Peak static overpressure (Δp)
- Peak dynamic pressure
- Overpressure impulse
- Dynamic pressure impulse (I_q)
- Vehicle overturn ($\Delta p I_q - K$)

Low Altitude Electromagnetic Pulse

- Vertical electric field
- Radial electric field
- Azimuthal magnetic field

An example of these formulas will be given in the Applications section. Once the distance D has been calculated as described above, the desired environmental criterion, E_c, can be determined by selecting the appropriate formula to compute the environment at range D. This environment, E_c, is then exceeded at the adjacent target at most the fraction 1-C of the time, and equipment-hardening decisions could then be made based on this criterion.

3. RESULTS

Tables 1 through 11 present the values of D corresponding to the following choices of input parameters:

Separation distance, S:	2, 3, 4, ..., 12 km
Standard deviation, σ:	100, 150, 200, ..., 950, 1000 m
Confidence level, C:	0.99, 0.95, 0.90, 0.75, 0.50

4. APPLICATIONS

4.1 Command Post Survivability

A subject of considerable current interest is the specification of the optimal command post architecture for the integrated battlefield. Studies have been made of the enhancement of the survivability of command posts through dispersion and redundancy measures.[2] In the dispersed command post structure, separate independent functions -- such as the air support operations center, the all-source intelligence center, the fire support elements, etc. -- are dispersed into separate areas, so that if one area happens to be destroyed, the other areas can continue to function. Although there are certainly other considerations, such as the extra communications needed which may limit the distance permitted between various elements, one

[2] John R. Bondanella, "Corps Command Post Architecture for the 1986-1990 Integrated Battlefield - A Vulnerability Analysis," thesis, U. S. Army Command and General Staff College, Ft. Leavenworth, Kansas (June 1980).

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 2.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	1.77	1.84	1.88	1.94	2.00
150	1.66	1.76	1.81	1.91	2.01
200	1.55	1.69	1.76	1.88	2.01
250	1.44	1.61	1.70	1.85	2.02
300	1.34	1.53	1.64	1.82	2.02
350	1.23	1.46	1.59	1.80	2.03
400	1.13	1.39	1.54	1.77	2.04
450	1.03	1.33	1.48	1.75	2.05
500	0.94	1.26	1.44	1.73	2.06
550	0.85	1.20	1.39	1.71	2.08
600	0.76	1.14	1.35	1.70	2.09
650	0.68	1.08	1.30	1.68	2.11
700	0.61	1.03	1.26	1.57	2.12
750	0.55	0.98	1.23	1.66	2.14
800	0.49	0.93	1.19	1.65	2.16
850	0.45	0.69	1.16	1.64	2.18
900	0.42	0.86	1.14	1.63	2.20
950	0.40	0.83	1.11	1.63	2.23
1000	0.38	0.80	1.09	1.62	2.25

Table 1.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 3.0 KM.

SIGMA (M)	CONFIDENCE LEVEL				
	.99	.95	.90	.75	.50
100	2.77	2.64	2.87	2.93	3.00
150	2.66	2.76	2.81	2.90	3.00
200	2.55	2.68	2.75	2.87	3.01
250	2.44	2.60	2.69	2.84	3.01
300	2.33	2.53	2.63	2.81	3.02
350	2.22	2.45	2.58	2.79	3.02
400	2.11	2.38	2.52	2.76	3.03
450	2.00	2.30	2.46	2.73	3.04
500	1.90	2.23	2.41	2.71	3.04
550	1.80	2.16	2.36	2.69	3.05
600	1.70	2.09	2.30	2.66	3.06
650	1.60	2.02	2.25	2.64	3.07
700	1.50	1.96	2.20	2.62	3.08
750	1.40	1.89	2.15	2.60	3.10
800	1.31	1.83	2.11	2.58	3.11
850	1.22	1.77	2.06	2.56	3.12
900	1.14	1.71	2.02	2.55	3.14
950	1.06	1.65	1.98	2.53	3.15
1000	0.98	1.59	1.94	2.52	3.17

Table 2.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 4.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	3.77	3.84	3.87	3.93	4.00
150	3.66	3.76	3.81	3.90	4.00
200	3.54	3.68	3.75	3.87	4.01
250	3.43	3.60	3.69	3.84	4.01
300	3.32	3.52	3.63	3.81	4.01
350	3.21	3.44	3.57	3.78	4.02
400	3.10	3.37	3.51	3.75	4.02
450	2.99	3.29	3.45	3.73	4.03
500	2.88	3.22	3.40	3.70	4.03
550	2.78	3.14	3.34	3.67	4.04
600	2.67	3.07	3.28	3.65	4.05
650	2.57	3.00	3.23	3.62	4.06
700	2.46	2.93	3.18	3.60	4.06
750	2.36	2.86	3.12	3.57	4.07
800	2.26	2.79	3.07	3.55	4.08
850	2.16	2.72	3.02	3.53	4.09
900	2.06	2.65	2.97	3.51	4.11
950	1.97	2.59	2.92	3.49	4.12
1000	1.87	2.52	2.87	3.47	4.13

Table 3.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 5.0 KM.

SIGMA	CONFIDENCE LEVEL				
(M)	.99	.95	.90	.75	.50
100	4.77	4.84	4.87	4.93	5.00
150	4.66	4.76	4.81	4.90	5.00
200	4.54	4.68	4.75	4.87	5.01
250	4.43	4.60	4.69	4.84	5.01
300	4.32	4.52	4.63	4.81	5.01
350	4.21	4.44	4.57	4.78	5.01
400	4.10	4.36	4.51	4.75	5.02
450	3.99	4.29	4.45	4.72	5.02
500	3.88	4.21	4.39	4.69	5.03
550	3.77	4.13	4.33	4.66	5.03
600	3.66	4.06	4.27	4.64	5.04
650	3.55	3.98	4.22	4.61	5.05
700	3.45	3.91	4.16	4.58	5.05
750	3.34	3.84	4.11	4.56	5.06
800	3.24	3.76	4.05	4.53	5.07
850	3.13	3.69	4.00	4.51	5.08
900	3.03	3.62	3.94	4.48	5.09
950	2.93	3.55	3.89	4.46	5.09
1000	2.83	3.43	3.84	4.44	5.10

Table 4.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 6.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	5.77	5.84	5.87	5.93	6.00
150	5.66	5.76	5.81	5.90	6.00
200	5.54	5.68	5.75	5.87	6.00
250	5.43	5.60	5.69	5.84	6.01
300	5.32	5.52	5.63	5.81	6.01
350	5.21	5.44	5.56	5.78	6.01
400	5.09	5.36	5.50	5.75	6.02
450	4.98	5.28	5.44	5.72	6.02
500	4.87	5.20	5.38	5.69	6.02
550	4.76	5.13	5.33	5.66	6.03
600	4.65	5.05	5.27	5.63	6.03
650	4.54	4.98	5.21	5.60	6.04
700	4.43	4.90	5.15	5.57	6.04
750	4.33	4.83	5.09	5.55	6.05
800	4.22	4.75	5.04	5.52	6.06
850	4.11	4.68	4.98	5.49	6.06
900	4.01	4.60	4.93	5.47	6.07
950	3.90	4.53	4.87	5.44	6.08
1000	3.80	4.46	4.82	5.42	6.09

Table 5.

Table of Distances in Kilometers for
Separation Distance of 7.0 km

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	6.77	6.84	6.87	6.93	7.00
150	6.66	6.76	6.81	6.90	7.00
200	6.54	6.68	6.75	6.87	7.00
250	6.43	6.60	6.69	6.84	7.01
300	6.32	6.52	6.62	6.81	7.01
350	6.20	6.44	6.56	6.77	7.01
400	6.08	6.36	6.50	6.74	7.01
450	5.98	6.28	6.44	6.71	7.02
500	5.87	6.20	6.38	6.68	7.02
550	5.76	6.12	6.32	6.65	7.02
600	5.65	6.05	6.26	6.62	7.03
650	5.54	5.97	6.20	6.60	7.03
700	5.43	5.89	6.15	6.57	7.04
750	5.32	5.82	6.09	6.54	7.04
800	5.21	5.74	6.03	6.51	7.05
850	5.10	5.67	5.97	6.48	7.06
900	4.99	5.59	5.92	6.46	7.06
950	4.88	5.52	5.86	6.43	7.07
1000	4.78	5.45	5.80	6.41	7.08

Table 6.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 8.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	7.77	7.84	7.87	7.93	8.00
150	7.66	7.76	7.81	7.90	8.00
200	7.54	7.68	7.75	7.87	8.00
250	7.43	7.60	7.69	7.84	8.01
300	7.32	7.52	7.62	7.80	8.01
350	7.20	7.44	7.56	7.77	8.01
400	7.09	7.36	7.50	7.74	8.01
450	6.98	7.28	7.44	7.71	8.02
500	6.87	7.20	7.38	7.68	8.02
550	6.75	7.12	7.32	7.65	8.02
600	6.64	7.04	7.26	7.62	8.03
650	6.53	6.97	7.20	7.59	8.03
700	6.42	6.89	7.14	7.56	8.03
750	6.31	6.81	7.08	7.53	8.04
800	6.20	6.74	7.02	7.51	8.04
850	6.09	6.66	6.96	7.48	8.05
900	5.99	6.58	6.91	7.45	8.06
950	5.88	6.51	6.85	7.42	8.06
1000	5.77	6.43	6.79	7.40	8.07

Table 7.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 9.0 KM.

SIGMA	CONFIDENCE LEVEL				
(M)	.99	.95	.90	.75	.50
100	8.77	8.84	8.87	8.93	9.00
150	8.66	8.76	8.81	8.90	9.00
200	8.54	8.68	8.75	8.87	9.00
250	8.43	8.59	8.68	8.84	9.00
300	8.31	8.51	8.62	8.80	9.01
350	8.20	8.43	8.56	8.77	9.01
400	8.09	8.36	8.50	8.74	9.01
450	7.98	8.28	8.44	8.71	9.01
500	7.86	8.20	8.38	8.68	9.02
550	7.75	8.12	8.32	8.65	9.02
600	7.64	8.04	8.26	8.62	9.02
650	7.53	7.96	8.20	8.59	9.03
700	7.42	7.88	8.14	8.56	9.03
750	7.31	7.81	8.08	8.53	9.04
800	7.20	7.73	8.02	8.50	9.04
850	7.09	7.65	7.96	8.47	9.04
900	6.98	7.58	7.90	8.44	9.05
950	6.87	7.50	7.84	8.42	9.05
1000	6.76	7.43	7.79	8.39	9.06

Table 8.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 10.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	9.77	9.84	9.87	9.93	10.00
150	9.66	9.76	9.81	9.90	10.00
200	9.54	9.68	9.75	9.87	10.00
250	9.43	9.59	9.68	9.84	10.00
300	9.31	9.51	9.62	9.80	10.01
350	9.20	9.43	9.56	9.77	10.01
400	9.09	9.35	9.50	9.74	10.01
450	8.98	9.27	9.44	9.71	10.01
500	8.86	9.20	9.38	9.68	10.02
550	8.75	9.12	9.31	9.65	10.02
600	8.64	9.04	9.25	9.62	10.02
650	8.53	8.96	9.19	9.59	10.02
700	8.42	8.88	9.13	9.56	10.03
750	8.30	8.80	9.07	9.53	10.03
800	8.19	8.73	9.01	9.50	10.04
850	8.08	8.65	8.95	9.47	10.04
900	7.97	8.57	8.90	9.44	10.05
950	7.86	8.50	8.84	9.41	10.05
1000	7.75	8.42	8.78	9.38	10.06

Table 9.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 11.0 KM.

SIGMA	CONFIDENCE LEVEL				
(M)	.99	.95	.90	.75	.50
100	10.77	10.84	10.87	10.93	11.00
150	10.66	10.75	10.81	10.90	11.00
200	10.54	10.67	10.75	10.87	11.00
250	10.45	10.59	10.63	10.84	11.00
300	10.31	10.51	10.62	10.80	11.01
350	10.20	10.43	10.56	10.77	11.01
400	10.09	10.35	10.50	10.74	11.01
450	9.97	10.27	10.44	10.71	11.01
500	9.86	10.19	10.37	10.68	11.01
550	9.75	10.12	10.31	10.65	11.02
600	9.64	10.04	10.25	10.62	11.02
650	9.52	9.96	10.19	10.58	11.02
700	9.41	9.88	10.13	10.55	11.04
750	9.30	9.80	10.07	10.52	11.03
800	9.19	9.72	10.01	10.49	11.03
850	9.08	9.65	9.95	10.46	11.04
900	8.97	9.57	9.89	10.44	11.04
950	8.86	9.49	9.83	10.41	11.05
1000	8.75	9.41	9.77	10.38	11.05

640

Table 10.

TABLE OF DISTANCES IN KM.
FOR SEPARATION DISTANCE OF 12.0 KM.

SIGMA (M)	CONFIDENCE LEVEL .99	.95	.90	.75	.50
100	11.77	11.84	11.87	11.93	12.00
150	11.66	11.76	11.81	11.90	12.00
200	11.54	11.67	11.75	11.87	12.00
250	11.43	11.59	11.68	11.84	12.00
300	11.31	11.51	11.62	11.80	12.01
350	11.20	11.43	11.56	11.77	12.01
400	11.09	11.35	11.50	11.74	12.01
450	10.97	11.27	11.44	11.71	12.01
500	10.86	11.19	11.37	11.68	12.01
550	10.75	11.11	11.31	11.64	12.02
600	10.63	11.03	11.25	11.61	12.02
650	10.52	10.96	11.19	11.58	12.02
700	10.41	10.88	11.13	11.55	12.02
750	10.30	10.80	11.07	11.52	12.03
800	10.19	10.72	11.01	11.49	12.03
850	10.08	10.64	10.95	11.46	12.03
900	9.97	10.56	10.89	11.43	12.04
950	9.85	10.49	10.83	11.40	12.04
1000	9.74	10.41	10.77	11.37	12.05

Table 11.

significant question is the dispersion required between the elements to alleviate bonus damage. It is this question that the algorithm in this report is well-suited to address. A similar question arises in the case of the desirable separation between redundant command posts in an architecture in which survivaiblity is increased by simply multiplying the number of command posts; this increase in survivability is then traded off with the cost of duplication.

The answer to the question of optimal separation depends on the environmental criteria pertinent to the units under study. For example, suppose the command posts are designed to withstand 2600 rad total radiation dose to personnel, 3 psi peak static overpressure, and 10^4 v/m EMP vertical electric field. (The effects of thermal on personnel are not considered since most personnel in a command post should be in a trailer or other protected structure.) By means of the environment formulas discussed in section 2, these criteria occur at a specific distance, D, from a weapon burst of a given yield. For example, the formula for peak static overpressure is given by:

$$\Delta p \text{ (psi)} = 1.61 \left(\frac{d}{w^{1/3}} \right)^{-1.70} \qquad (3)$$

for a weapon of yield w in KT at a scaled height of burst of 60 $w^{1/3}$ m. Other formulas may be found in the cited reference.[1] Suppose the likely threats to the command posts are 300- and 600-kT weapons delivered with a total standard deviation, σ, of 500 m, and that bonus damage is to be precluded with a confidence C = 0.95. The environmental criteria then correspond to the ranges, D, shown in Table 12; and interpolation between the appropriate Tables 1 through 11 with D as the search variable produces the separation distances, S, shown.

	w = 300 kT		w = 600 kT	
	D	S	D	S
Criterion	(km)	(km)	(km)	(km)
2600 rad	1.9	2.7	2.1	2.9
3 psi	4.6	5.2	5.9	6.7
10^4 v/m	3.2	4.0	3.6	4.2

Table 12. Separation distance calculations

As can be seen, overpressure is the dominant of the environments considered; and a separation distance of at least 6.7 km between the command post elements or the redundant command posts is required to preclude bonus damage from the assumed threat with 95% confidence.

4.2 Weapon Employment

In this example Blue targeteers are planning an attack strategy against particular Red radars integral to air defense units which are typically located only 2 to 2.5 km apart. The Blue weapon to be used against the radars is a 50-kT missile which is delivered at two-thirds maximum range with a standard deviation, σ, of 250 m. The dominant kill mechanism to the radars is blast and the peak static overpressure damage criterion is 2.4 psi. This overpressure level occurs at a distance, D, of 2.91 km from such a weapon burst.

For a 50% level of confidence, interpolation in Tables 1 through 11 gives a separation distance, S, of 2.89 km, as that distance at which the probability is 50% that bonus damage at the adjacent target will exceed the overpressure criterion. Since the air defense units are pairwise separated by less than 2.89 km, there is a greater than 50% probability that firing a round at one of the Red radars will also produce blast damage to the radar located in the neighbor unit.

4.3 Hardening/Operational Deployment Trade-off

This example examines the effectiveness trade-off between the operational separation maintained between Army units of a certain type and the hardening of the equipment in these units to various levels of nuclear blast. Suppose the likely Red threat to this type of unit is a 600 kT weapon delivered with a total standard deviation, σ, of 500 m. The dominant kill mechanism to this unit is taken to be peak static overpressure, Δp, to various critical pieces of equipment in the unit.

Three levels of hardening of this equipment will be considered:

- 1 psi (sure-safe criterion)
- 2 psi
- 5.8 psi (man-survivability criterion)

The distances, D, at which these levels of Δp are encountered are calculated from equation (3) and presented in Table 13. If it is desired to preclude bonus damage to adjacent units of the same type at a confidence level of 95%, then interpolation in Tables 1 through 11 yields the separation distances, S, which must be maintained between the units, as shown in Table 13.

Blast Criterion (psi)	D (km)	S (km)
1	11.2	12.0
2	7.4	8.2
5.8	4.0	4.8

Table 13. Hardening level vs. separation distance

Consequently if the equipment in these units is vulnerable to blast levels of 1 psi, then the units must be located at least 12 km apart to ensure that there is no more than a 5% probability that a round aimed at one such unit could not only destroy that unit but also inflict bonus damage on the neighbor unit.

On the other hand, if the units are hardened to the man-survivability level of 5.8 psi, then they can be deployed as close as 4.8 km apart and be protected from bonus damage (with 95% confidence).

5. CONCLUSIONS

In conclusion, this paper has presented an algorithm for assessing bonus damage which accounts for weapon-delivery errors and target-location error, and attaches confidence bounds to the results derived. This methodology is applicable to a wide spectrum of problems, examples of which have been provided; the reader should be able to discover other appropriate problems from his own experience.

A GENERALIZED RAYLEIGH-RITZ METHOD FOR STRUCTURAL DYNAMICS PROBLEMS IN CONJUNCTION WITH FINITE ELEMENTS

Julian J. Wu
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. A solution formulation of generalized Rayleigh-Ritz method is described and applied to two initial and boundary value problems of stress waves and structural dynamics in conjunction with finite element discretization. Excellent numerical results have been obtained for wave equations associated with lateral and longitudinal vibrations and with strong discontinuities.

I. INTRODUCTION. This paper describes a solution formulation for and its applications to initial boundary value problems of structural dynamics and stress waves. Excellent numerical results are stated in conjunction with finite element discretization. The basic concept of this approach is to establish a variational problem equivalent to a given initial boundary value problem, which is in general, non-self-adjoint, through the use of an adjoint field variable and the use of some large "spring" constants so that all the end conditions can be transformed into natural "boundary" conditions. Therefore, the shape functions used need not satisfy any end conditions a priori in solving the variational problem in the same manner as applying the Rayleigh-Ritz method for self-adjoint problems. This same concept was demonstrated in solving initial value problems in a paper delivered at the International Symposium on Numerical Methods in Engineering Science series in 1978 and later published in the Journal of Sound and Vibration [1]. In this present paper, the formulation is extended to initial boundary value problems and the numerical results obtained are also encouraging.

In the section which follows immediately, two initial boundary value problems are stated. One is a longitudinal stress wave problem in a rod. There is a discontinuity in the initial data given. We wish to trace this discontinuity in the numerical solution using the present approach. The second problem is a beam vibration problem under a moving concentrated load. This is a much more difficult problem since the partial differential equation is of fourth order and the force is singular in nature. In the next section, variational problems equivalent to the given initial boundary problems are established. The finite element discretization procedures are then briefly recaptured. Lastly, numerical results are presented with some comments.

II. INITIAL BOUNDARY VALUE PROBLEMS. Two initial boundary problems of structural dynamics will be stated in this Section. The first one is of longitudinal elastic stress wave in a rod with a sudden change in initial conditions. The second one is concerned with lateral vibrations of a Euler-Bernoulli beam subjected to a moving concentrated load.

Longitudinal Stress Wave in a Rod. The rod is fixed at one end and free at the other end. The discontinuity data arises from the initial linear displacement, corresponding to a constant stress, due to a force applied at the "free" end. This force suddenly disappears at time zero causing a stress discontinuity at the free end. The differential equation can be written as:

$$\frac{{}^2u}{\partial x^2} = \frac{1}{a^2}\frac{\partial^2 u}{\partial t^2} \quad ; \quad \begin{array}{l} 0 < x < \ell \\ 0 < t < T \end{array} \tag{1}$$

with

$$a^2 = E/\rho \tag{2}$$

where u = u(x,t) is the axial displacement; x,t are the coordinates in axial direction and in time, respectively; ρ,E are density and Young's modulus, respectively, of the rod material; ℓ denotes length of the rod; and T denotes some finite time of interest.

For the boundary conditions, we have

$$u(0,t) = 0$$

and

$$\frac{\partial u}{\partial x}(1,t) = 0 \tag{3}$$

The dynamics of the problem are due to the initial conditions. It is assumed that the rod is stretched to a linear displacement by a force P which vanishes at time t > 0 (see Figure 1). The initial velocity of the rod is assumed to be zero. Thus

$$u(x,0) = \frac{P}{AE}\,x \quad ; \quad \text{and}$$

and

$$\frac{\partial u}{\partial t}(x,0) = 0 \tag{4}$$

It is convenient to use dimensionless parameters. Let

$$u^* = u/\ell \quad , \quad x^* = x/\ell \quad , \quad t^* = t/T \tag{5}$$

Then, Eq. (1) in dimensionless form is

$$\frac{\partial^2 u^*}{\partial x^{*2}} = b^2\,\frac{\partial^2 u^*}{\partial t^{*2}} \quad , \quad \begin{array}{l} 0 < x^* < 1 \\ 0 < t^* < 1 \end{array} \tag{6}$$

where

$$b^2 = \frac{1}{a^2} \left(\frac{\ell}{T}\right) \tag{7}$$

The boundary conditions become

$$u^*(0,t^*) = 0 \quad , \quad \frac{\partial u^*}{\partial x^*} (1,t^*) = 0 \tag{8}$$

and

$$u^*(x,0) = P^*x^* \quad ; \quad \frac{\partial u^*}{\partial t^*} (x^*,0) = 0 \tag{9}$$

where

$$P^* = \frac{P}{AE} \tag{10}$$

is the force in dimensionless form.

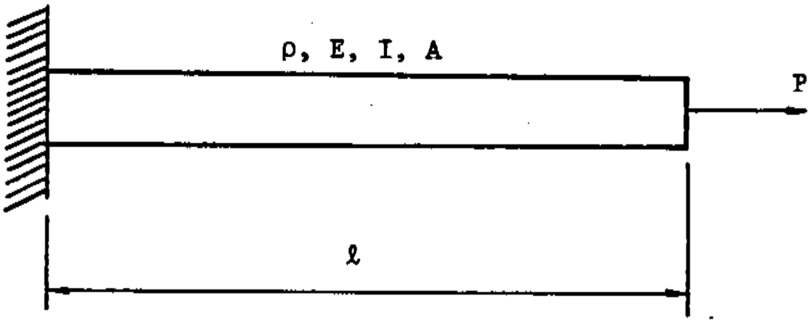


Figure 1. A Rod Fixed at One End and Subjected to a Load P, which is Suddenly Released at Time Zero.

The stated problem in dimensionless form combines Eqs. (6), (8), and (9) with the new dimensionless parameters related to physical counterparts by Eqs. (5), (6), and (10). To simplify writing, we shall drop the asterisks (*) in Eqs. (6), (8), and (9), and rewrite them as

$$u'' - b^2\ddot{u} = 0 \quad ; \quad \begin{matrix} 0 \leq x \leq 1 \\ 0 \leq t \leq 1 \end{matrix} \tag{6'}$$

$$u(0,t) = 0 \quad ; \quad u'(1,t) = 0 \tag{8'}$$

$$u(x,t) = Px \quad ; \quad \dot{u}(x,0) = 0 \tag{9'}$$

where a prime (') indicates differentiation with respect to x and a dot (•), with respect to t.

Beam Vibrations Under Moving Loads. Let us consider the differential equation of a uniform Euler-Bernoulli beam subjected to a moving, concentrated force.

$$EIy'''' + \rho A\ddot{y} = \bar{P}\delta(x_p - x) \qquad \begin{matrix} 0 \leq x \leq \ell \\ 0 \leq t \leq T \end{matrix} \tag{11}$$

where

E,ρ	=	Young's modulus, density of the beam material
I,A	=	second moment, area of the beam's cross-section
ℓ	=	length of the beam
$y=y(x,t)$	=	beam deflection
x,t	=	coordinates in beam's axial direction and in time
$\underline{P}$	=	magnitude of the concentrated force
$\delta(x)$	=	Dirac delta function
$x_p=x_p(t)$	=	location of P
T	=	some finite time of interest

Again it will be convenient to employ nondimensional parameters and equations. These will be introduced by way of Eq. (11). Thus, let

$$y^* = y/\ell \quad , \quad x^* = x/\ell \quad , \quad t^* = t/T \tag{12}$$

Using Eq. (12) in Eq. (11), one has

$$y^{*''''} + \gamma^2 y^* = \bar{Q}\delta(x_p^* - x^*) \qquad \begin{matrix} 0 \leq x^* \leq 1 \\ 0 \leq x^* \leq 1 \end{matrix} \tag{13}$$

where

$$\gamma = \frac{c}{T} \quad , \quad c^2 = \frac{\rho A\ell^4}{EI} \quad , \quad Q = P^* = \frac{P\ell^2}{EI} \tag{14}$$

Also note in Eq. (13) that the differentiations are now with respect to the nondimensionalized variables x* ant t*. From now on, we shall use Eq. (13) with the asterisks dropped altogether.

$$y'''' + ky + \gamma^2\ddot{y} = \bar{Q}\delta(x_p-x) \qquad \begin{matrix} 0 < x < 1 \\ 0 < t < 1 \end{matrix} \tag{15}$$

III. VARIATIONAL PROBLEMS - A GENERALIZED RAYLEIGH-RITZ METHOD. For the stress wave problem in the previous section, consider a variational problem.

$$\delta I_o = 0 \tag{16a}$$

with

$$I_o = I_o(u,v) = \int_0^1 \int_0^1 (-u'v'+b^2\dot{u}\dot{v})dxdt \tag{16b}$$

where u(x,t) and v(x,t) are said to be adjoint to each other. It is a simple matter to see that this problem is an indeterminate one. However, the functional of Eq. (16b) can be modified to a variational problem which is equivalent to the boundary/initial problem of Eqs. (6'), (8'), and (9'). Thus consider

$$\delta I = 0 \tag{17a}$$

with

$$\begin{aligned} I = I(u,v) = & \int_0^1 \int_0^1 (-u'v'+b^2\dot{u}\dot{v})dxdt \\ & + k_1\int_0^1 u(0,t)v(0,t)dt \\ & + k_2b^2\int_0^1 [u(x,0) - u_o(x)]v(x,1)dx + b^2 \int_0^1 u_1(x)v(x,0)dx \end{aligned} \tag{17b}$$

We shall take the first variation of the function I(u,v) of Eq. (17b) in such a manner that δv is completely arbitrary while δu is set to zero identically. Hence, by means of integration-by-parts, one has

$$\begin{aligned} (\delta I)_{\delta u=0} = & \int_0^1 \int_0^1 (u''-b^2\ddot{u})\delta v dxdt \\ & - \int_0^1 u'(1,t)\delta v(1,t)dt \\ & + \int_0^1 [u(0,t) + k_1u(0,t)]\delta v(0,t)dt \\ & + b^2\int_0^1 \{\dot{u}(x,1) + k_2[u(x,0) - u_o(x)]\}\delta v(x,1)dx \\ & - b^2 \int_0^1 [\dot{u}(x,0) - u_1(x)]\delta v(x,0)dx = 0 \end{aligned} \tag{18}$$

The fact that δv(x,t) is completely arbitrary enables us to conclude from Eq. (18) that

$$u'' - b^2 u = 0 \quad ; \quad \begin{matrix} 0 < x < 1 \\ 0 < t < 1 \end{matrix} \tag{19a}$$

$$u'(1,t) = 0$$

$$u'(0,t) + k_1 u(0,t) = 0$$

$$u(x,1) + k_2[u(x,0) - u_0(x)] = 0 \tag{19b}$$

and

$$u(x,0) - u_1(x) = 0$$

It is then observed that the initial boundary value problem defined by Eqs. (19a) and (19b) reduces to that of Eqs. (6'), (8'), and (9') if one lets k_1 and k_2 go to infinity* (and with $u_0(x) = Px$ and $u_1(x) = 0$). This fact suggests that the variational problem of Eqs. (17a) and (17b) can be used as a basis of a finite element discretization for the approximate solutions to the original initial boundary problem. It should be noted that all the auxiliary conditions in Eqs. (19a) and (19b) are the so-called natural boundary conditions. They are the consequence of the variational problem - just like the differential equation itself. For this reason, the above solution is referred to as a Generalized Rayleigh-Ritz Method.

By a similar process, one can establish a variational problem for the vibration problem of a beam under a moving load. In this case, one has

$$\delta I = \int_0^1 \int_0^1 [u''\delta v'' - \dot{u}\delta\dot{v} - \bar{\delta}(x-\bar{x})\delta v]dxdt$$

$$+ \int_0^1 [k_1 u(0,t)\delta v(0,t) + k_2 u'(0,t)\delta v'(0,t)$$

$$+ k_3 u(1,t)\delta v(1,t) + k_4 u'(1,t)\delta v'(1,t)]dt$$

$$+ \int_0^1 [k_5 u(x,0)\delta v(x,1) + k_6 \dot{u}(x,0)\delta v(x,0)]dx = 0 \tag{20}$$

*This process is sometines referred to as the penalty function method. See, for example, Reference [2].

Through integrations-by-parts,

$$\delta I = \int_0^1 \int_0^1 [u'''' + \ddot{u} - \bar{\delta}(x-\bar{x})]\delta v(x,t)dxdt$$

$$+ \int_0^1 \{[k_1 u(0,t) + u'''(0,t)]\delta v(0,t) + [k_2 u'(0,t) - u''(0,t)]\delta v'(0,t)$$

$$+ [k_3 u(1,t) - u'''(1,t)]\delta v(1,t) + [k_4 u'(1,t) + u''(1,t)]\delta v'(1,t)\}dt$$

$$+ \int_0^1 \{[k_5(u(x,0)-0) - \dot{u}(x,t)]\delta v(x,1) + (k_6+1)[\dot{u}(x,0)-0]\delta v(x,0)\}dx = 0 \quad (21)$$

The original differential equation and the boundary and initial conditions are recovered from the equation above due to the arbitrariness of the variations $\delta(x,t)$ and by properly selecting the values of k_i,s, $i = 1,2,...,6$.

IV. FINITE ELEMENT DISCRETIZATION. Only essential features will be stated in the finite element discretizations here. The region of a unit square ($0 \leq x \leq 1$; $0 \leq t \leq 1$) is further divided into KxL rectangles by taking K divisions in x direction and L divisions in t direction. Local coordinates (ξ,η) in each (i,j)th element are related to (x,t) by these equations:

$$\begin{aligned} \xi &= \xi^{(i)} = Kx - i + 1 \\ \eta &= \eta^{(j)} = Lt - j + 1 \end{aligned} \qquad (22)$$

Within each element, the unknown function u(x,t) is replaced by the approximation:

$$\begin{aligned} u_{(i,j)}(\xi,\eta) &= \underset{\sim}{a}^T(\xi,\eta)\ U_{(i,j)} \\ \delta v_{(i,j)}(\xi,\eta) &= \underset{\sim}{a}^T(\xi,\eta)\ \delta V_{(i,j)} \end{aligned} \qquad (23)$$

where $a(\xi,\eta)$ is the shape function vector and $U_{(i,j)}$, $\delta V_{(i,j)}$ are the generalized coordinates. The specific form of $a(\xi,\eta)$ employed here is such that each one of the sixteen components is:

$$a_k(\xi,\eta) = b_i(\xi)b_j(\eta) \quad , \quad \begin{array}{l} k = 1,2,.....,16 \\ i,j = 1,2,3,4 \end{array} \qquad (24)$$

with

$$\begin{aligned} b_1(\varepsilon) &= 1 - 3\xi^2 + 2\xi^3 \\ b_2(\xi) &= \xi - 2\xi^2 + \xi^3 \\ b_3(\xi) &= 3\xi^2 - 2\xi^3 \\ b_4(\xi) &= -\xi^2 + \xi^3 \end{aligned} \qquad (25)$$

and the relations between index k and the pair (i,j) are given in Table I.

TABLE I. RELATIONSHIP BETWEEN (i,j) AND k IN EQUATION (24)

k	(i,j)	k	(i,j)
1	(1,1)	9	(1,3)
2	(2,1)	10	(2,3)
3	(1,2)	11	(1,4)
4	(2,2)	12	(2,4)
5	(3,1)	13	(3,3)
6	(4,1)	14	(4,3)
7	(3,2)	15	(3,4)
8	(4,2)	16	(4,4)

Using Eqs. (22) through (25) in Eq. (17) and the fact that $V_{(i,j)}$ is completely arbitrary, the matrix equations for the unknowns $U_{(i,j)}$ can be routinely assembled and solved. Further details will be omitted here.

V. NUMERICAL RESULTS AND DISCUSSION. Some of the numerical results are presented in this section. For the stress wave problem*, Table II provides solutions of v(x,t), $\partial u/\partial x(x,t)$ and $\partial u/\partial t(x,t)$ for x = 0, 0.1, 0.2, ...1.0 and for t = 0, 0.5, 1.0, 1.5, and 2.0. During this time interval, the original displacement has gone through a complete sign reversal as shown in Figure 2. This particular set of data was obtained by taking K = 10 and L = 1 with restart procedures, i.e., the final solution in the first time step was taken as the initial condition of the next step in time, and so on. Values of the exact solutions are given in parentheses. Excellent agreement is observed. The fact that the discontinuity of the solution follows along without much oscillation is worth mentioning.

For the beam vibration problem with a moving force, some typical numerical solutions are given in Tables III and IV. The moving concentrated force is assumed to travel at a constant velocity c (although this is not at all a restriction for the present method) such that

$$x(t) = ct$$

where c is dimensionless velocity. For small c, c = 0.0001, and the displacement solutions become those of static deflections as shown in Table III. For a large c (compared with unity), c = 10, and solutions show dynamic effects as indicated in Table IV. As a comparison, solutions obtained by the Fourier series and Laplace transform method [4] are given in parentheses. Good agreement exists even in cases with considerable dynamic effect.

*For exact solution to this problem, see for example, Reference [3].

TABLE II. SOLUTIONS TO THE STRESS WAVE PROBLEM OF EQS. (6'), (8') and (9') WITH b = 1.0, P = 1.0.
(PART 1)

	Data at Time t = 0.0			Data at Time t = 0.50			Data at Time t = 1.00		
x	u(x,t)	∂u/∂x	∂u/∂t	u(x,t)	∂u/∂x	∂u/∂t	u(x,t)	∂u/∂x	∂u/∂t
0.0	0.00000 (0.00000)	1.00000 (1.00000)	0.00000 (0.00000)	-0.00000 (0.00000)	0.99861 (1.00000)	0.00000 (0.00000)	0.00000 (0.00000)	-0.11718 (0.00000)	0.00000 (0.00000)
0.10	0.10000 (0.10000)	1.00000 (1.00000)	0.00000 (0.00000)	0.09998 (0.10000)	0.99740 (1.00000)	-0.00142 (0.00000)	-0.01236 (0.00000)	0.18416 (0.00000)	-1.07661 (-1.00000)
0.20	0.20000 (0.20000)	1.00000 (1.00000)	0.00000 (0.00000)	0.19994 (0.20000)	0.99113 (1.00000)	-0.00618 (0.00000)	0.00259 (0.00000)	-0.10702 (0.00000)	-1.09024 (-1.00000)
0.30	0.30000 (0.30000)	1.00000 (1.00000)	0.00000 (0.00000)	0.29949 (0.30000)	0.97359 (1.00000)	-0.01886 (0.00000)	-0.00026 (0.00000)	-0.00077 (0.00000)	-0.92271 (-1.00000)
0.40	0.40000 (0.40000)	1.00000 (1.00000)	0.00000 (0.00000)	0.40354 (0.40000)	1.06038 (1.00000)	0.07965 (0.00000)	0.00035 (0.00000)	-0.00953 (0.00000)	-1.03476 (-1.00000)
0.50	0.50000 (0.50000)	1.00000 (1.00000)	0.00000 (0.00000)	0.49976 (0.50000)	0.47813 (1.00000)	-0.54638 (0.00000)	-0.00036 (0.00000)	-0.00287 (0.00000)	-0.95721 (-1.00000)
0.60	0.60000 (0.60000)	1.00000 (1.00000)	0.00000 (0.00000)	0.49785 (0.50000)	0.02932 (0.00000)	-0.96310 (-1.00000)	0.00042 (0.00000)	-0.00334 (0.00000)	-1.04064 (-1.00000)
0.70	0.70000 (0.70000)	1.00000 (1.00000)	0.00000 (0.00000)	0.50081 (0.50000)	-0.02913 (0.00000)	-1.06490 (-1.00000)	-0.00039 (0.00000)	0.00046 (0.00000)	-0.95841 (-1.00000)
0.80	0.80000 (0.80000)	1.00000 (1.00000)	0.00000 (0.00000)	0.49983 (0.50000)	0.00258 (0.00000)	-0.95849 (-1.00000)	0.00041 (0.00000)	-0.00086 (0.00000)	-1.04185 (-1.00000)
0.90	0.90000 (0.90000)	1.00000 (1.00000)	0.00000 (0.00000)	0.50019 (0.50000)	0.00036 (0.00000)	-1.03927 (-1.00000)	-0.00040 (0.00000)	-0.00012 (0.00000)	-0.95859 (-1.00000)
1.00	1.00000 (1.00000)	0.00000 (0.00000)	0.00000 (0.00000)	0.49982 (0.50000)	-0.00000 (0.00000)	-0.96235 (-1.00000)	0.00041 (0.00000)	0.00000 (0.00000)	-1.04164 (-1.00000)

*Figures in parentheses indicate exact solutions.

TABLE II. SOLUTIONS TO THE STRESS WAVE PROBLEM OF EQS. (6'), (8'), and (9') with b = 1.0, P = 1.0.
(PART 2)

	Data at Time t = 1.50			Data at Time t = 2.00		
x	u(x,t)	∂u/∂x	∂u/∂t	u(x,t)	∂u/∂x	∂u/∂t
0.0	0.00000 (0.00000)	-0.99528 (-1.00000)	0.00000 (0.00000)	0.00000 (0.00000)	-0.02019 (-1.00000)	-0.00000 (0.00000)
0.10	-0.9975 (-0.10000)	-1.01418 (-1.00000)	0.00235 (0.00000)	-0.10003 (-0.10000)	-0.98040 (-1.00000)	0.00165 (0.00000)
0.20	-0.19988 (-0.20000)	-0.93268 (-1.00000)	-0.04972 (0.00000)	-0.19992 (-0.20000)	-0.01911 (-1.00000)	-0.00119 (0.00000)
0.30	-0.30222 (-0.30000)	-1.13452 (-1.00000)	0.09812 (0.00000)	-0.30008 (-0.30000)	-0.97758 (-1.00000)	0.00610 (0.00000)
0.40	-0.39324 (-0.40000)	-0.62190 (-1.00000)	-0.02522 (0.00000)	-0.39981 (-0.40000)	-0.00984 (-1.00000)	-0.00308 (0.00000)
0.50	-0.49607 (-0.50000)	0.19515 (0.00000)	-0.40290 (0.00000)	-0.50035 (-0.50000)	-0.98060 (-1.00000)	-0.02259 (0.00000)
0.60	-0.51154 (-0.50000)	0.00219 (0.00000)	-1.14247 (-1.00000)	-0.59945 (-0.60000)	-1.01522 (-1.00000)	0.02001 (0.00000)
0.70	-0.49849 (-0.50000)	-0.00219 (0.00000)	-1.05659 (-1.00000)	-0.69929 (-0.70000)	0.84502 (-1.00000)	-0.05514 (0.00000)
0.80	-0.50034 (-0.50000)	-0.02908 (0.00000)	-0.92639 (-1.00000)	-0.80180 (-0.80000)	-0.84502 (-1.00000)	0.07546 (0.00000)
0.90	-0.49935 (-0.50000)	-0.01126 (0.00000)	-1.03406 (-1.00000)	-0.90529 (-0.90000)	-1.19080 (-1.00000)	-0.23643 (0.00000)
1.00	-0.50051 (-0.50000)	-0.00000 (0.00000)	-0.93441 (-1.00000)	-0.98802 (-1.0000)	-0.00000 (-1.00000)	0.31500 (0.00000)

*Figures in parentheses indicate exact solutions.

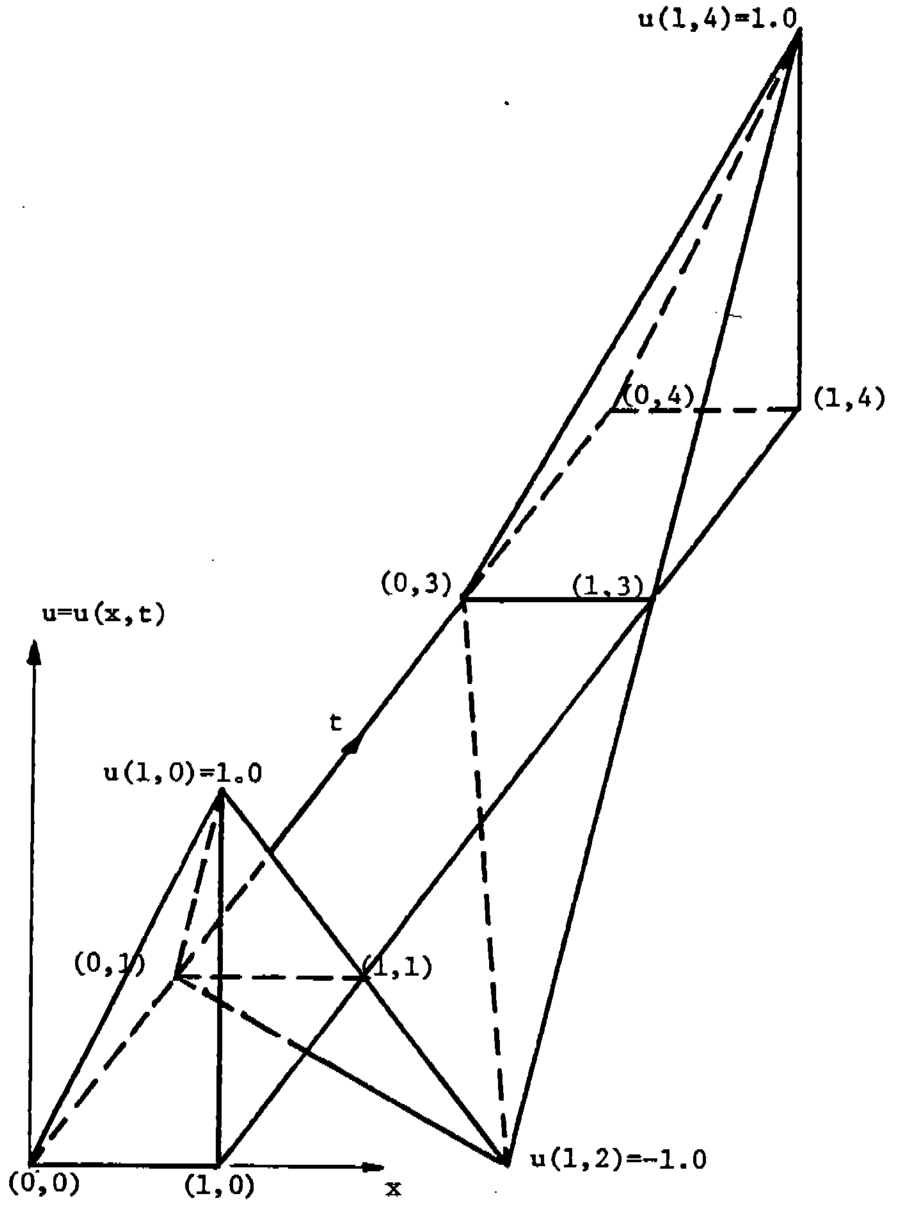


Figure 2. Exact Solution Surface u=u(x,t) for the Stress Wave Problem of Eqs. (6'), (8'), and (9') in the Region: $0 \leqslant x \leqslant 1$ and $0 \leqslant t \leqslant 4$ (with $P = 1.0$ and $b = 1.0$).

TABLE III. SOLUTIONS u(x,t) TO THE MOVING FORCE PROBLEM OF EQ. (15) WITH Q = 1.0 AND FIXED END CONDITIONS AT x = 0. (For very low velocity, $\gamma = 10^{-4}$ in Eq. (14))

x	t = 0.0	t = 0.20	t = 0.40	t = 0.60	t = 0.80	t = 1.00
0.0	0.000000 (Given)	0.000000 (0.000000)	0.000000 (0.000000)	0.000000 (0.000000)	0.000000 (0.000000)	-0.000000 (0.000000)
0.20	0.000001 (Given)	0.008533 (0.009534)	0.011999 (0.012000)	0.010665 (0.010667)	0.006125 (0.006133)	-0.000097 (0.000000)
0.40	-0.000001 (Given)	0.012001 (0.012000)	0.019206 (0.019199)	0.018157 (0.018134)	0.010763 (0.010666)	0.001174 (0.000000)
0.60	-0.000000 (Given)	0.010668 (0.010667)	0.018137 (0.018133)	0.019214 (0.019201)	0.012057 (0.012000)	0.000691 (0.000000)
0.80	0.000001 (Given)	0.006133 (0.006134)	0.010664 (0.010666)	0.011990 (0.012000)	0.008491 (0.008533)	-0.000509 (0.000000)
1.00	-0.000000 (Given)	0.000000 (0.000000)	0.000000 (0.000000)	0.000000 (0.000000)	0.000000 (0.000000)	0.000000 (0.000000)

Solutions in parentheses based on formulas from Reference 4.

TABLE IV. SOLUTIONS u(x,t) TO THE MOVING FORCE PROBLEM OF EQ. (15) WITH Q = 1.0 AND FIXED END CONDITIONS AT x = 0. (For very low velocity, γ = 10 in Eq. (14))

x	t = 0.0	t = 0.20	t = 0.40	t = 0.60	t = 0.80	t = 1.00
0.0	-0.000000 (Given)	-0.000000 (0.000000)	-0.000000 (0.000000)	-0.000000 (0.000000)	-0.000000 (0.000000)	-0.000000 (0.000000)
0.20	-0.000001 (Given)	0.000482 (0.000467)	0.001387 (0.001345)	0.002151 (0.002046)	0.002944 (0.002534)	0.007191 (0.003643)
0.40	-0.000001 (Given)	-0.000077 (-0.000082)	0.001109 (0.001109)	0.002717 (0.002704)	0.004364 (0.004375)	0.004877 (0.004463)
0.60	-0.000000 (Given)	0.000001 (0.000025)	-0.000320 (-0.000311)	0.001493 (0.001504)	0.003110 (0.003177)	0.005067 (0.005601)
0.80	-0.000001 (Given)	0.000003 (-0.000012)	-0.000013 (0.000002)	-0.000964 (-0.000942)	0.001308 (0.001257)	0.005751 (0.005464)
1.00	-0.000000 (Given)	-0.000000 (0.000000)	0.000000 (0.000000)	-0.000000 -(0.000000)	-0.000000 (0.000000)	-0.000000 (0.000000)

Solutions in parentheses based on formulas from Reference 4.

In conclusion, this paper has demonstrated through examples of structural dynamics an approximate solution formulation (which is both a weighted method and a variational problem), the finite element implementation, and some favorable numerical results. Although only linear problems have been mentioned, an application to solutions of non-linear problems is now being investigated.

REFERENCES

1. J. J. Wu, "Solutions to Initial Value Problems by Use of Finite Elements-Unconstrained Variational Formulations," 1977 Journal of Sound and Vibration, 53, pp. 341-356.

2. D. G. Luenberger, Optimization by Vector Space Method, John Wiley, 1969, p. 302.

3. L. S. Jacobson and R. S. Ayre, Engineering Vibrations, McGraw-Hill, 1965, pp. 472-474.

4. L. Fryba, Vibrations of Solids and Structures Under Moving Load, Noordhoff, 1971.

A Numerical Technique in the Hydrodynamic Theory of Foil Bearings

I.G. Tadjbakhsh*, G. Ahmadi** and E.A. Saibel+

Abstract

The hydrodynamic theory of foil bearings is reviewed. The relationship between fluid pressure and film thickness is discussed. The compressibility of gas is included in the analysis. It is shown that the basic equation for determination of pressure distribution becomes a third order boundary value problem in terms of film thickness. A simple numerical scheme for solution of the nonlinear boundary value problem is developed and some examples are considered and discussed.

*Department of Civil Engineering, Rensselaer Polytechnic Institute, Troy, New York.

**Department of Mechanical and Industrial Engineering Clarkson College, Potsdam, New York.

+U.S. Army Research Office, Research Triangle Park, North Carolina.

1. Introduction

Theory of hydrodynamic lubrication is one of the well established fields of mechanical engineering [1,2]. The theory of hydrodynamic lubrication with deformable boundaries was considered by Korovchinskii [3], Christensen [4,5], Wilson [6] and Mahdariaw and Wilson [7], among others.

The theory of foil bearing was first investigated by Blok and Van Rossum [8]. Wildman and Wright [9] have considered the effect of external pressure on foil bearings and have also employed a perturbation method for solution of the resulting equations. Further developments are carried out by Eshel and Elrod [10], Ma [11], Barlow [12] and more recently by Eshel [13,14].

In the present investigation, the theory of gas lubrication with a flexible boundary is studied. The equations of motion of a flexible tape are considered and under the assumption of small slope a simple relationship between the fluid-pressure and film thickness is established. The basic equation for the variation of film pressure is then obtained which is shown to be a nonlinear third order two point boundary value problem. The general expressions for the load bearing capacity and the friction force of the bearing are derived and discussed. A simplified scheme for obtaining the numerical solution of the formulated nonlinear boundary value problem is developed and applied to several examples.

2. Basic Equations

We consider a hydrodynamical bearing which is driven by a moving flexible tape or belt as shown in Fig. 1. As a result of the pressure developed within the lubricating film the flexible tape deflects from the straight line position joining the two roller bearings at the entrance and exit of the bearing.

We assume that the steadily moving tape assumes a fixed shape in space, i.e., the tape slides along a fixed curve c in the x-y plane. Let 0 be a fixed point in the space located also on the curve of the moving belt. Let 0' be a point on the moving tape serving as a reference point for measurement of the distance S of the material points along the tape and which coincides with the point 0 at a reference time t_o. A point P a distance S away from 0' along curve c at time t_o will be a distance $\xi = S + v(t-t_o)$ away from 0 at time t. The Cartesian coordinates $\bar{x}$ and $\bar{y}$ of the point P as well as the tension τ of the tape at that point will be functions of the variable ξ. The equations of motion of the tape, assuming no resistance to bending, and inextensibility condition are:

$$\frac{\partial}{\partial S}\left(\tau \frac{\partial \bar{x}}{\partial S}\right) + f_x = \gamma \frac{\partial^2 \bar{x}}{\partial t^2}, \tag{1}$$

$$\frac{\partial}{\partial S}\left(\tau \frac{\partial \bar{y}}{\partial S}\right) + f_y = \gamma \frac{\partial^2 \bar{y}}{\partial t^2}, \tag{2}$$

$$\left(\frac{\partial \bar{x}}{\partial S}\right)^2 + \left(\frac{\partial \bar{y}}{\partial S}\right)^2 = 1. \tag{3}$$

Here f_x and f_y are the fluid forces in x and y directions exerted on the tape and γ is the mass per unit length of the tape.

Since bearings generally have small slopes, it can be said that $\frac{\partial \bar{y}}{\partial \xi} = \frac{\partial \bar{y}}{\partial S} << 1$ and hence $\frac{\partial \bar{x}}{\partial \xi} = \frac{\partial \bar{x}}{\partial S} = 1$. Using this and the fact that $\bar{x}$, $\bar{y}$ and τ are functions of ξ, (1) and (2) can be written as

$$\frac{d\tau}{d\xi} + f_x = \gamma\, v^2 \; \frac{d^2\bar{x}}{d\xi^2} \approx 0 \,, \tag{4}$$

$$f_y = \gamma\, v^2 \; \frac{d^2\bar{y}}{d\xi^2}. \tag{5}$$

Under the stated set of assumptions, we have

$$f_x = 2\mu \, \frac{\partial u}{\partial y} \Big|_{y=h} \,, \tag{6}$$

$$f_y = - p \,, \tag{7}$$

in which μ is the viscosity of the lubricant and u and p represent the fluid velocity and excess (gage) pressure as governed by the Reynolds equation

$$\frac{\partial^2 u}{\partial y^2} = \frac{1}{\mu}\, \frac{dp}{dx} \,. \tag{8}$$

Integrating equation (8) and using the boundary conditions

$$u=0 \text{ at } y=0 \quad , \tag{9}$$

$$u = \frac{\partial \bar{x}}{\partial t} \approx v \text{ at } y = h \,, \tag{10}$$

we obtain

$$u = \frac{1}{2\mu} \, \frac{dp}{dx} \, (y^2 - hy) + \frac{y}{h}\, v. \tag{11}$$

The continuity equation is given by

$$\rho_a Q = \int_0^h \rho u \, dy, \tag{12}$$

where ρ_a and ρ are the densities of gas (air) at atmospheric pressure P_a and at gage pressure p, respectively. For a polytropic precess, we have

$$\frac{p_a + p}{\rho^n} = \frac{p_a}{\rho_a^n} = \text{const.} \tag{13}$$

n equal to 1.4 and 1.0 correspond to the adiabatic and isothermal processes, respectively and n = ∞ denotes the isochoric (incompressible) gas flow. Using equation (11) in (12) and noting that ρ is only a function of x, it follows that

$$Q = (\frac{\rho}{\rho_a})(\frac{1}{2}\,vh - \frac{h^3}{12\mu}\;\frac{dp}{dx}). \tag{14}$$

Eliminating ρ between equation (13) and (14), we find

$$Q = (1 + \frac{p}{p_a})^{1/n}(\frac{1}{2}\,vh - \frac{h^3}{12\mu}\;\frac{dp}{dx}). \tag{15}$$

The linearized forms of equations (5) and (7) lead to

$$p = -\gamma v^2\,\frac{d^2h}{dx^2}\;. \tag{16}$$

Substituting equation (16) into (15), we find

$$(1 - \frac{\gamma v^2}{p_a} \frac{d^2h}{dx^2})^{1/n} (\frac{\gamma v^2}{12\mu} h^3 \frac{d^3h}{dx^3} + \frac{1}{2} vh) = Q. \tag{17}$$

For the incompressible limit, that is, $n = \infty$, equation (17) reduces, to that of [8,9]. The boundary conditions

$$h = h_o \quad \text{at} \quad x = 0 \ , \tag{18}$$

$$h = h_1 \quad \text{at} \quad x = \ell \ , \tag{19}$$

$$\frac{d^2h}{dx^2} = 0 \quad \text{at} \quad x = 0, \ell \ , \tag{20}$$

serve to determine the flow rate Q and the three constants of integration.

Introducing dimensionless variables,

$$T = \frac{h}{h_o} \ , \quad \eta = \frac{x}{\ell} \ , \tag{21}$$

equation (17) and boundary conditions (18) - (20) become

$$(1-\lambda\frac{d^2T}{d\eta^2})^{1/n} (T^3 \frac{d^3T}{d\eta^3} + \alpha T) = \beta, \quad 0 < \eta < 1 , \tag{22}$$

with

$$T = 1 \quad \text{at} \quad \eta = 0 \ , \tag{23}$$

$$T = \delta \quad \text{at} \quad \eta = 1 \ , \tag{24}$$

$$\frac{d^2T}{d\eta^2} = 0 \ \text{at} \quad \eta = 0,1 \ , \tag{25}$$

where

$$\alpha = \frac{6\mu\ell^3}{\gamma v h_o^3} \ , \ \beta = \frac{12\mu Q\ell^3}{\gamma v^2 h_o^4} \ , \quad \delta = \frac{h_1}{h_o} \ , \ \lambda = \frac{\gamma v^2 h_o}{p_a \ell^2} \tag{26}$$

parameter λ is the ratio of dynamic pressure to atmospheric pressure. For small values of λ, the gas behaves as an incompressible fluid.

Equations (22) - (25) form a nonlinear third order two point boundary value problem for finding the dimensionless film thickness T. The expression for the film pressure in terms of the dimensionless quantities is given by

$$p = - \frac{\gamma v^2 h_o}{\ell^2} \frac{d^2 T}{d\eta^2} \qquad (27)$$

The load bearing capacity per unit width is defined by

$$P = \int_o^\ell pdx \ . \qquad (28)$$

Employing equation (27) in (28), the expression for the load bearing capacity in terms of dimensionless quantities becomes

$$P = \frac{\gamma v^2 h_o}{\ell} [\frac{dT}{d\eta} (0) - \frac{dT}{d\eta} (1)] \ . \qquad (29)$$

From equations (29) it is observed that the bearing capacity is proportional to the square of tape velocity in contrast to the case of rigid boundaries where it becomes proportional directly to the velocity. Furthermore P is related to, h_o/ℓ, while in the rigid case it is proportional to ℓ^2/h_o. It is of course recognized that the terms $dT/d\eta$ in equation (29) depend on the values of parameters α, β and λ which are functions of v, h_o, ℓ, μ and etc. Therefore, the dependence of the load capacity on various parameters would be partially modified, accordingly.

The friction force per unit width of the bearing can be calculated by first obtaining the shear stress at the upper boundary and then integrating the result over the length ℓ. The final expression for the frictional force D becomes

$$D = \frac{2\mu v\ell}{h_o} \int_0^1 \frac{2T-3\beta/2\alpha}{T^2} d\eta \qquad (30)$$

From equation (30) it is observed that the dependence of D on the parameters μ, v, ℓ and h_o is similar to that of a bearing with rigid boundaries.

3. Numerical Solution and Examples

The numerical solution of the third order nonlinear two point boundary value problem given by equations (22) - (25) becomes rather involved The values of the parameter β (which depends on Q) and the slope $\frac{dT}{d\eta}(0)$ must be guessed in such a way that at $\eta = 1$ the values of $T = 1$ and $\frac{d^2T}{d\eta^2} = 0$ be reached simultaneously. Such a procedure requires a time consuming trial and error calculation. To circumvent the lengthy computation the following scheme for solution is adopted. Let us assume that

$$\frac{dT}{d\eta}(0) = k . \tag{31}$$

Introducing a change of independent variable,

$$r = k\eta , \tag{32}$$

equation (22) becomes

$$(1 - \lambda' \frac{d^2T}{dr^2})^{1/n} (T^3 \frac{d^3T}{dr^3} + \alpha' T) = \beta' \qquad 0 < r < k \tag{33}$$

where

$$\alpha' = k^{-3}\alpha, \quad \beta' = k^{-3}\beta, \; \lambda' = k^2\lambda. \tag{34}$$

The boundary conditions (23) - (25) and equation (31) now become,

$$T = 1, \; \frac{dT}{dr} = 1, \; \frac{d^2T}{dr^2} = 0, \qquad \text{at } r = 0 , \tag{35}$$

$$T = \delta, \; \frac{d^2T}{dr^2} = 0 , \qquad \text{at } r = k , \tag{36}$$

For fixed values of α', β' and λ' (assumed values) equation (33) can be integrated numerically with initial values of T, dT/dr and d^2T/dr^2 given by equation (35). Integration is carried out until d^2T/dr^2 becomes equal to zero. The corresponding r for which $d^2T/dr^2 = 0$ determines the value of the parameter k and the value of T at that position gives the magnitude of the parameter δ. Q can then be obtained from β'. This technique, can be used to generate a series of solutions for various values fo the parameters. The method is quite simple and avoids cumbersome trial and error procedure.

Introducing the change of variable (32) into equations (27) and (29) the expressions for the pressure and the load bearing capacity, respectively, become

$$p = -\frac{\gamma^2 v^2 h_o k^2}{\ell^2} \frac{d^2T}{dr^2}, \tag{37}$$

$$P = \frac{\gamma v^2 h_o k}{\ell} [1 - \frac{dT}{dr}(k)]. \tag{38}$$

A Runge-Kutta numerical integration scheme is employed and examples are considered. In the first example, the values of parameters are taken to be $\alpha' = 1$, and $\beta' = 0.667$ and $\lambda' = 0.0$ (that correspond to an incompressible fluid and a flow rate of $Q = \beta' v h_o / 2\alpha' = v h_o/3$). The numerical solution of equation (33) with initial conditions given by equation (35) is obtained and it is observed at $r = k = 6.22$, d^2T/dr^2 becomes approximately equal to zero. The value of parameters α, β and Q as found from (34) thus become

$$\alpha = 240.6, \quad \beta = 160.4, \quad Q = v h_o/3. \tag{39}$$

The corresponding variation of h/h_o with η is shown in figure 2. It is observed that the film thickness increases up to a maximum of about three times of the entrance film thickness and then decreases to about 0.25 h_o at the exit.

The variation of the dimensionless pressure

$$\bar{p} = \frac{p\ell^2}{\gamma^2 v^2 h_o^2 k^2} , \tag{40}$$

is shown in figure 3. It is observed that $\bar{p}$ reaches a peak of about 0.9 very close to the exit. The load bearing capacity of the bearing is found to be

$$P = 18.69 \frac{\gamma v^2 h_o}{\ell} . \tag{41}$$

In the second example it is assumed that $\alpha' = 1.6$, $\beta' = 1.067$ and $\lambda' = 0.0$. The numerical solution yields

$$k = 4.64, \ \delta = 0.26 , \tag{42}$$

It then follows that

$$\alpha = 159.9, \quad \beta = 106.6, \quad Q = vh_o/3 \tag{43}$$

The variations of film thickness and dimensionless film pressure with η are shown in figures 4 and 5. The load bearing capacity now is found to be

$$P = 14.72 \frac{\gamma v^2 h_o}{\ell} . \tag{44}$$

It is observed that the load bearing capacity decreases with an increase of α (i.e., an increase in gas viscosity). As a third example we consider a compressible gas film with parameter values $\alpha' = 1.0$, $\beta' = 0.5$, $\lambda' = 0.1$ and $n = 1.4$. The solution of the initial value problem yields $k = 5.5450$ and $T_r(k) = -2.0884$. This results in a load bearing capacity

$$P = 17.191 \frac{\gamma v^2 h_o}{\ell} \tag{45}$$

The film thickness and pressure variations are shown in figures 2-7 and are comparable with those obtained by a perturbation method in [9].

4. Concluding Remarks

In the present work, hydrodynamic theory of foil bearing is considered and the effects on gas compressibility are studied. It is observed that the behavior of a foil bearing is quite different from that of a regular rigid boundary type. For instance, the film pressure distribution has a sharp peak near the exit of a foil bearing in contrast to the relatively smooth peak observed about the middle of the conventional rigid boundary types. Furthermore, the dependence of the load bearing capacity of these flexible bearings is drastically different from the conventional one.

Several important problems such as optimization of the load bearing capacity, two dimensional effects, etc. are not treated here and are left to future investigations.

5. Acknowledgment

The support of the U. S. Army Research Office is gratefully acknowledged.

References

1. Pinkus, O. and Sternlicht, B., Theory of Hydrodynamic Lubrication, McGraw Hill, 1961.

2. O'Connor, J.J., Boyd, J. and Avallone, E.A., Standard Handbook of Lubrication Engineering, McGraw Hill, 1968.

3. Korovchinskii, M.V., "Some Problems in the Hydrodynamic Theory of Lubrication with Deformation of the Bodies Bounding the Lubricant Film", Third All Union Conference on Friction and Wear, Moscow, 1960.

4. Christenson, H., "The Oil Film in a Closing Gap", Proc. Roy. Soc. Series A, Vol. 266, 1962, p. 312.

5. Christenson, H., "Elasto-hydrodynamic Theory of Spherical Bodies in Normal Approach", Journal of Lubrication Technology, Transections of the ASME, Vol. 92, 1970, p. 145.

6. Wilson, W.R.D., "Film Thickness Variation in the Work Zone of Hydrodynamically Lubricated Continuous Deformation Processes", Journal of Lubrication Technology, Vol. 95, No. 4, 1973, pp. 541-543.

7. Mahdavian, S.M. and Wilson, W.R.D., "Lubricant Flow in a Plastohydrodynamic Work Zone", Journal of Lubrication Technology, Vol. 97, No. 1, 1976, pp. 16-21.

8. Blok, H., and Van Rossum, J.J., "The Foil Bearing - A New Departure in Hydrodynamic Lubrication", Lubrication Engineering, Vol. 9, No. 6, 1976, pp. 310-320.

9. Wildman, M. and Wright, A., "The Effect of External Pressurization on Self-Acting Foil Bearings", J. Basic Engineering, Trans. ASME Vol. 87, No. 3, 1965, pp. 631-640.

10. Eshel, A. and Elrod, Jr., M.G., "The Theory of the Infinitely Wide Perfectly Flexible, Self-Acting Foil Bearing", J. Basic Engineering, Trans. ASME, Vol. 87, No. 4, 1965, pp. 831-836.

11. Ma, J.T.S., "An Investigation of Self Acting Foil Bearings", J. Basic Engineering, Trans. ASME, Vol. 87, No. 4, 1965, pp. 837-846.

12. Barlow, E.J., "Externally Pressurized Foil Gas Bearings", J. Basic Engineering, Trans. ASME, Vol. 87, No. 4, 1965, pp. 986-990.

13. Eshel, A., "Reduction of Air Films in Magnetic Recording by External Air Pressure", J. Lubrication Technology, Trans. ASME, Vol. 96, No. 2, 1974, pp. 247-249.

14. Eshel, A., "Transient Analysis of a Planner Hybrid Foil Bearing Model", J. Lubrication Technology, Trans. ASME, Vol. 96, No. 3, 1974, pp. 432-436.

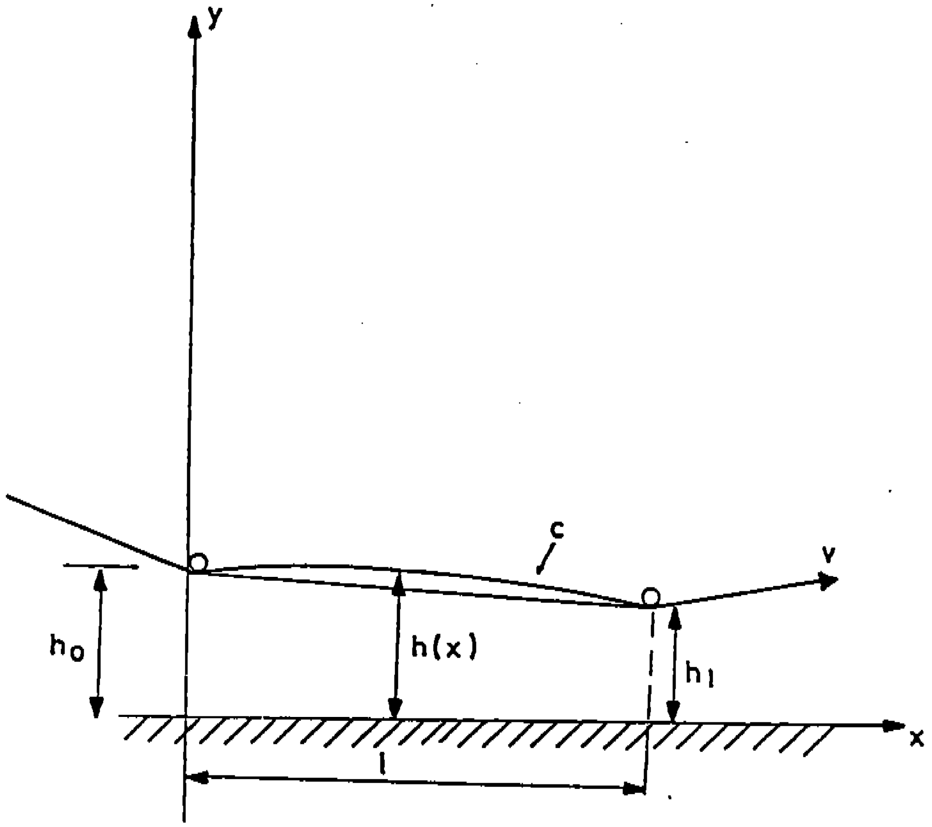


Fig. 1 - Sketch of the bearing.

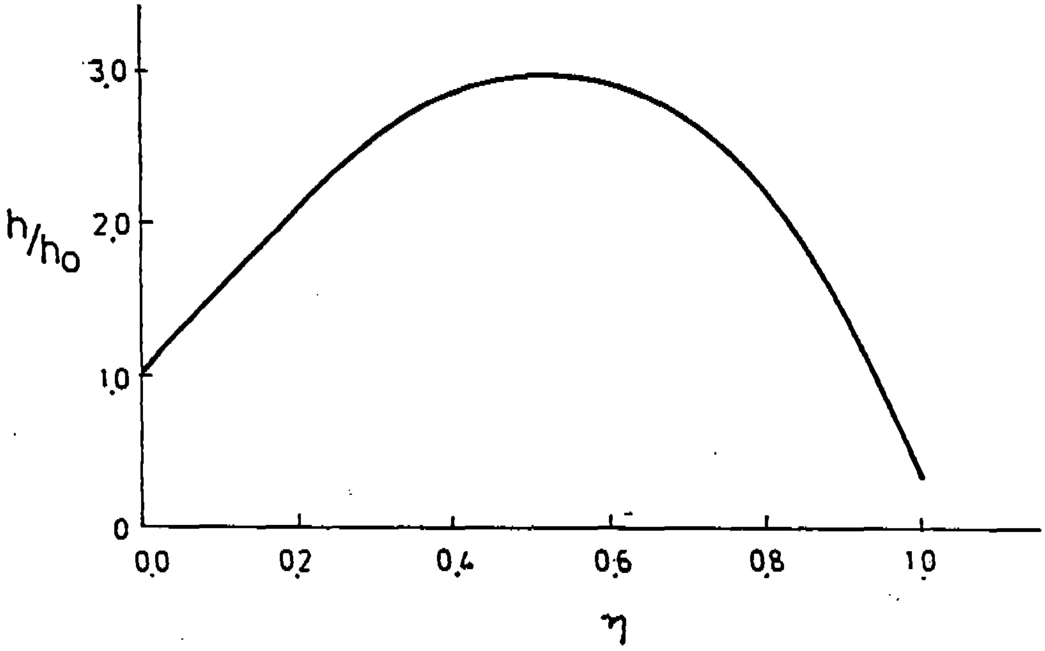


Fig. 2 - Variation of film thickness with η for α = 240.6, β = 160.4.

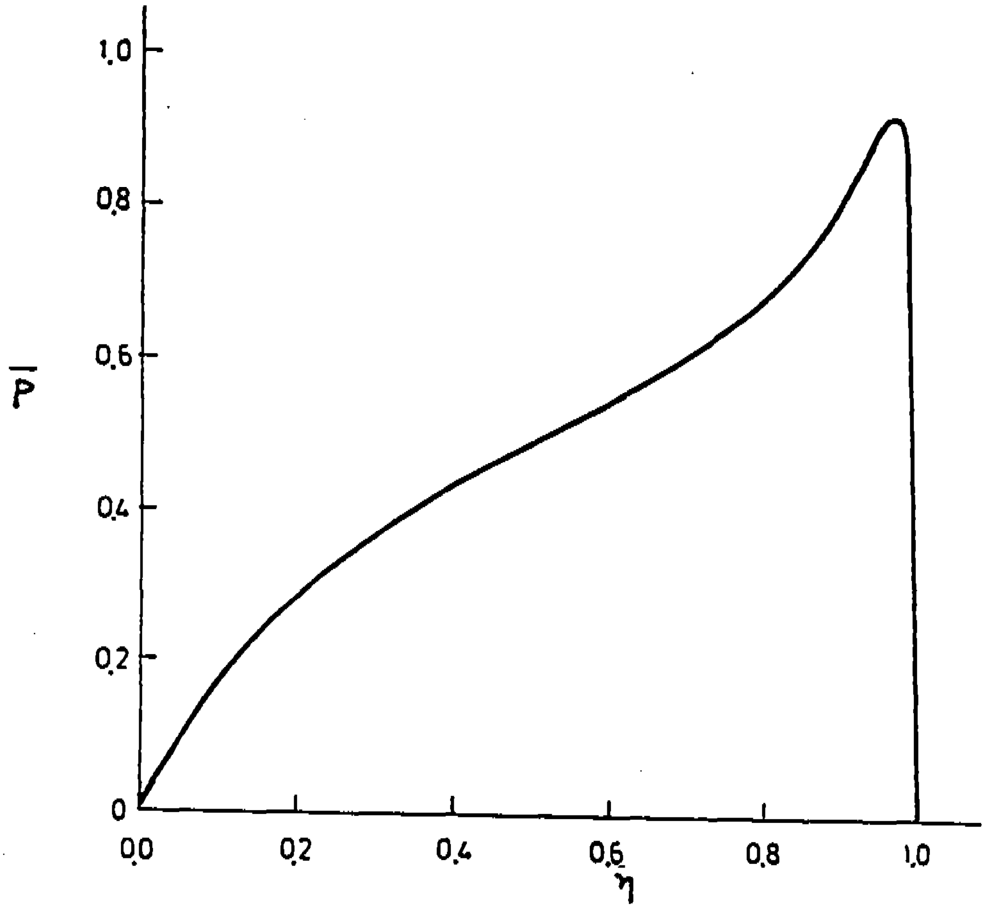


Fig. 3 - Variation of dimensionless film pressure with η for $\alpha = 240.6$, $\beta = 160.4$.

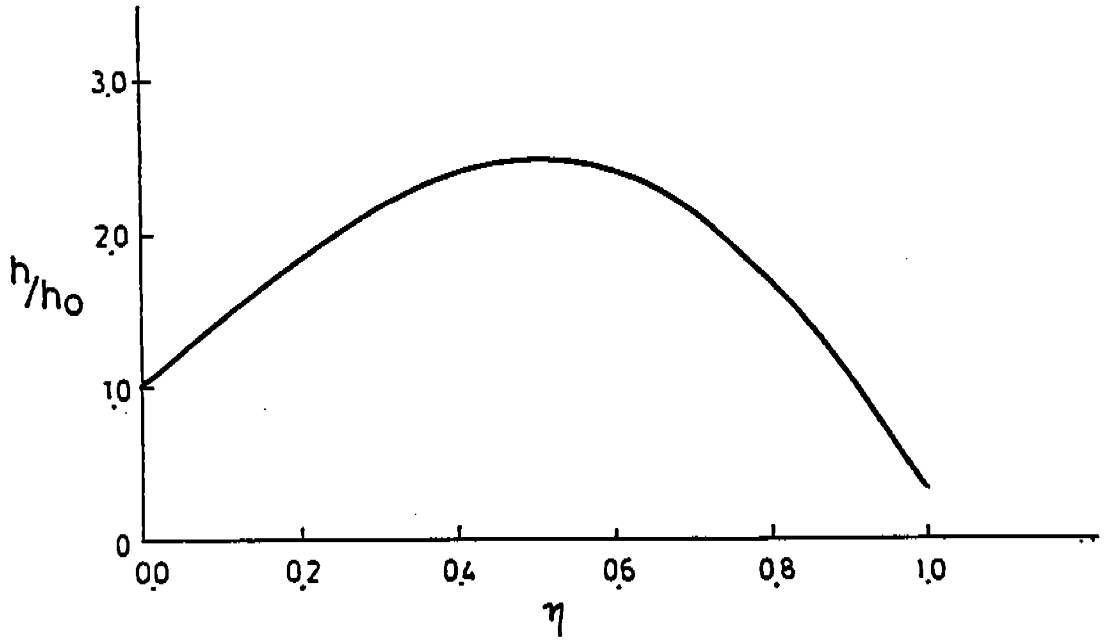


Fig. 4 - Variation of film thickness with η for $\alpha = 159.9$, $\beta = 106.6$.

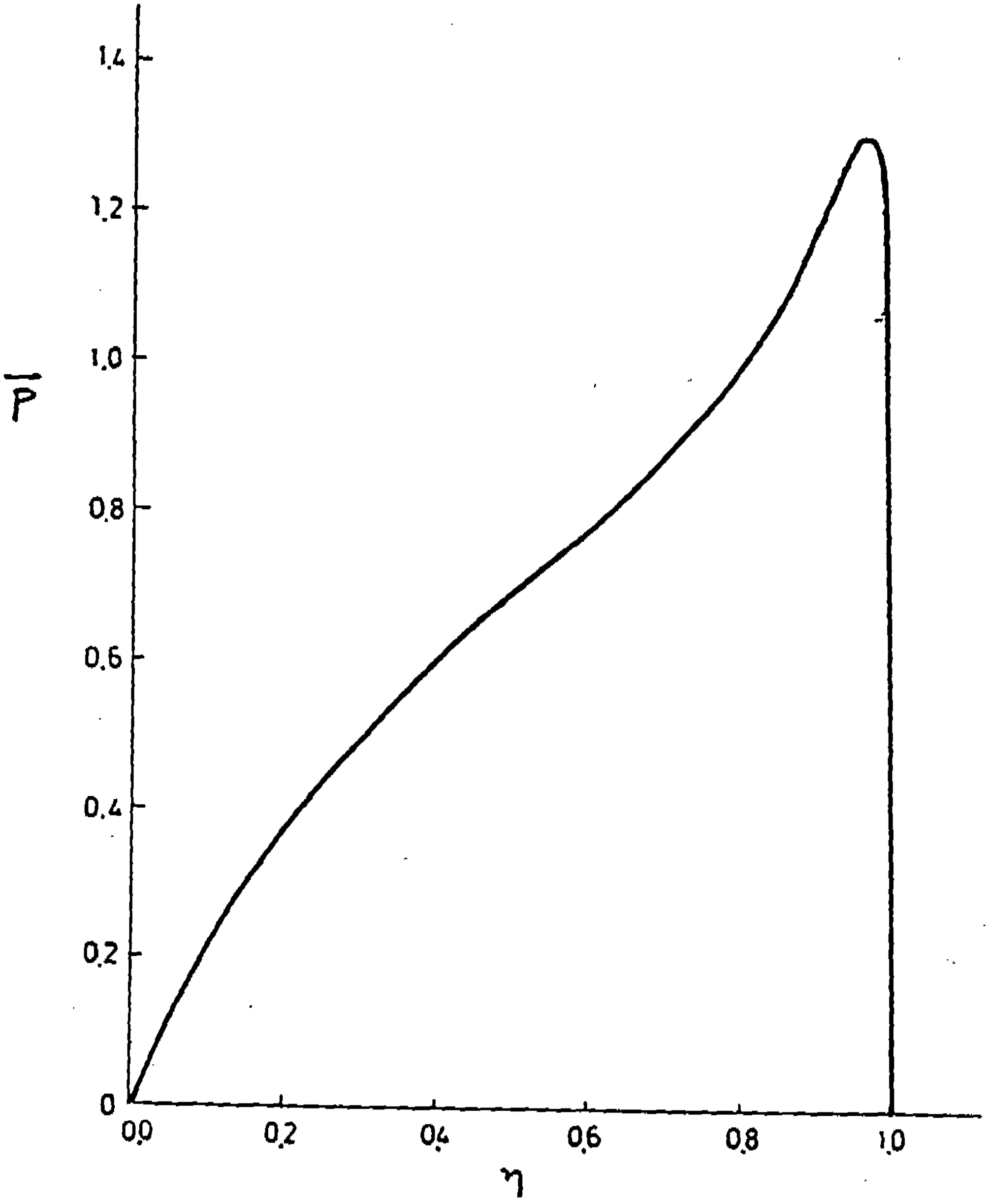


Fig. 5 - Variation of dimensionless film pressure with η for α = 159.9, β = 106.6 .

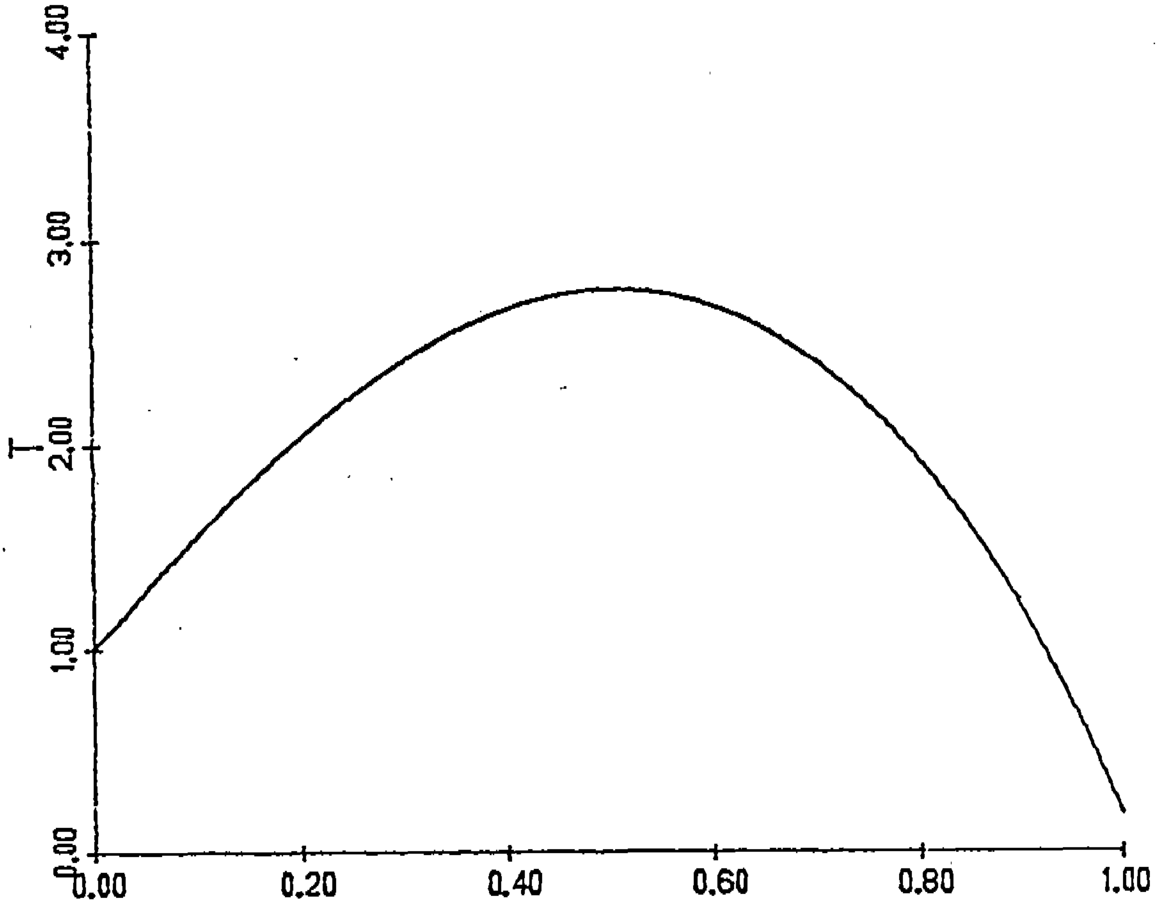


Figure 6 - Variation of film thickness with η for α = 170.5, β = 85.2 and λ = .0032.

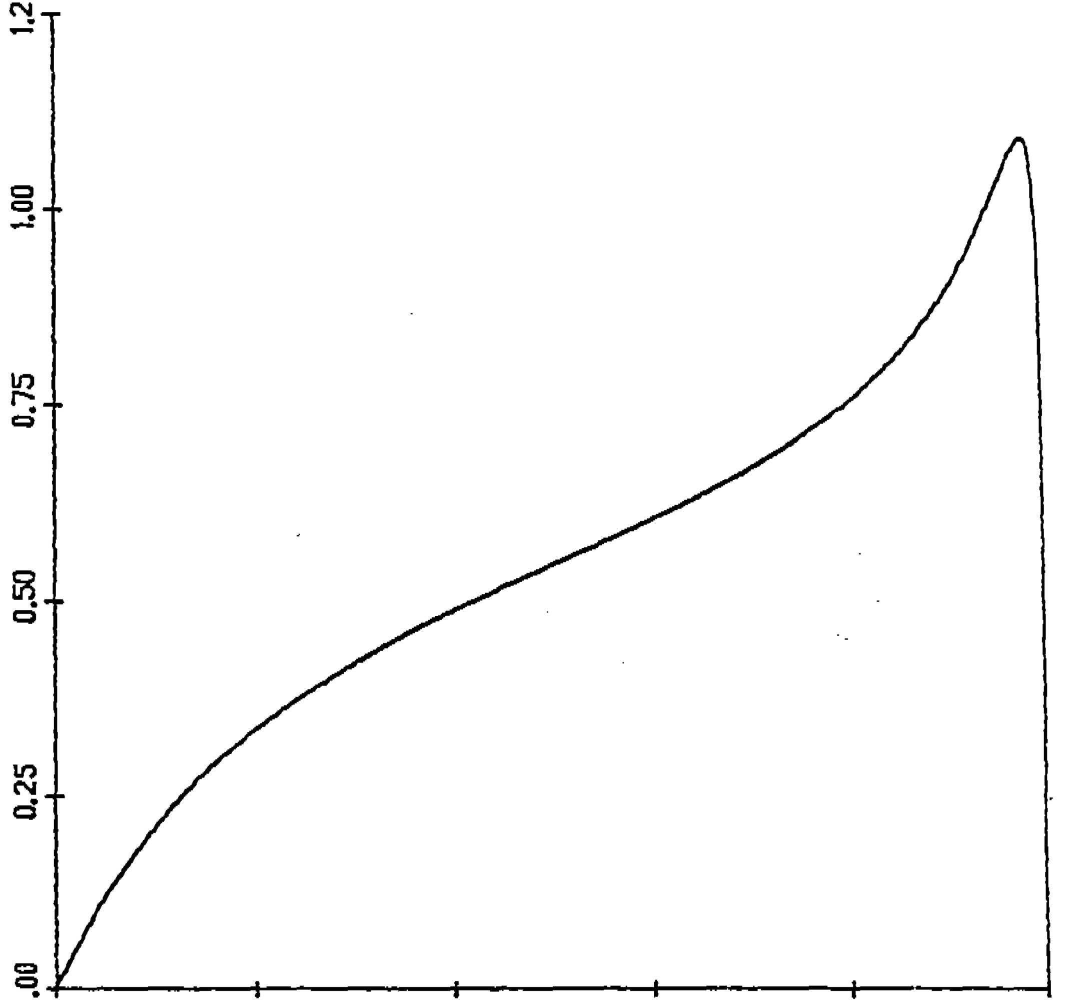


Figure 7 - Variation of dimensionless film pressure with η for α = 170.5, β = 85.2, λ = .0032.

INTERACTIVE COMBAT SIMULATION AS AN EXPERIMENTAL TOOL FOR THE EVOLUTION OF C^3I SYSTEMS

Reiner K. Huber

Institut für Angewandte Systemforschung
und Operations Research
Fachbereich Informatik
Hochschule der Bundeswehr München
D-8012 Neubiberg, West Germany

ABSTRACT: In the past two decades significant amounts of resources have been spent on the improvement and development of Command, Control, Communication, and Intelligence Systems (C^3I). The bureaucratic institutionalization of the cost-effectiveness approach within the Planning, Programming, and Budgeting (PPB) framework has contributed to the isolated assessment of C^3I resulting in highly centralized systems that maximize the use of modern technology for its own sake. However, C^3I is but one of five interdependent principal elements of defence systems requiring its assessment in terms of how much it contributes, together with the other elements, to the accomplishment of the overall defence mission. Hereby, the complementarity and the degree of substitutability between the physical and the conceptual elements of defence systems must be explicitly considered, not the least with a view to the limitation of available resources. Such an assessment may only be accomplished through dynamic analyses which account explicitly for the interaction of all system elements in combat. Thus, combat simulation becomes the principal tool of analysis. The incorporation of the military decision-maker into such simulations permits to also make the decision-making process object of exploratory research and to develop, in an evolutionary manner, decision support software. This is illustrated by two examples. One shows how cognitive maps of military commanders may be retrieved by means of interactive combat simulations. The other shows an approach to develop tactical decision models by means of such simulations. Given a certain technical capability, such research might be carried on within the framework of military exercises as proposed in the *Compound Gaming Approach*.

1. C^3I: AN ELEMENT OF MILITARY SYSTEMS

Command, Control, Communication and Intelligence (C^3I) has been defined as "... an arrangement of personnel, facilities and systems for information acquisition, processing and dissemination employed by a (military) decision-maker in planning, directing and controlling operations" (see [1],p. 42). Thus, even though it is not a readily separable entity such as a weapon system or a combat service support system, the C^3I system may be considered as one of the three fundamental physical elements of military systems competing for resources. For the force planner the question is how to distribute the available human and financial resources among these elements so that, in the prevailing threat environment, there is a high probability that the military missions essential to meeting the national objectives can be accomplished. Since the answer also depends on the operational philosophies and concepts of the respective military forces, we may, in a somewhat simplified manner,conceptualize the overall force planning problem in the form of a system of interacting physical and conceptual elements as depicted in Fig. 1.

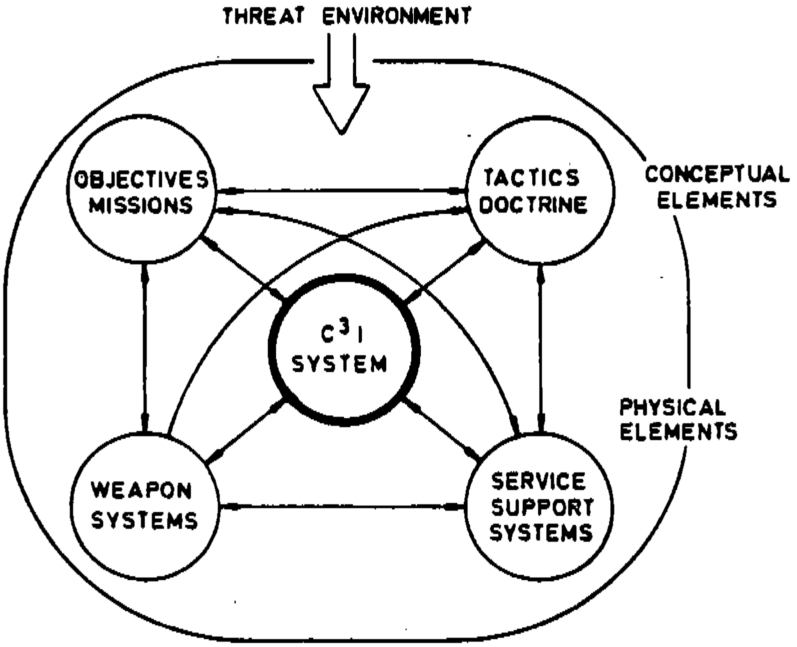


Fig. 1: Fundamental Elements of a Military Defence System

The C^3I system interacts with all the other force elements, not the least with the conceptual element of tactics and doctrine. For example, a military system employing the *Auftragstaktik* practiced in the German Armed Forces may very likely require comparatively little C^3. Rather than giving more or less detailed orders to the unit entrusted with a mission, according to *Auftragstaktik* merely the objectives are stated,while it is left to that unit's commander to determine how to accomplish the objectives. Thus, it is proposed that, when designing a military force and its C^3I system, also tactics and doctrine need to be reviewed and adapted to maximize the probability of mission accomplishment within the constraints imposed by the available resources.

2. THE PROBLEM OF PIECEMEAL ANALYSIS

It appears that reality of today's force planning hardly accounts for these interactions. Rather it assumes the conceptual elements as given parameters and considers each of the physical elements more or less in isolation. For example, force design analyses are mostly restricted to "optimizing" the weapons mix of military defence systems tacitly assuming the existence of support systems and C^3I capabilities that permit the employment of the weapon systems at their specified (maximum) performance level[1]. Tactics, doctrine and operational concepts are usually not analysis subjects but are considered invariant parameters (see Canby [4]). The *a priori* specified missions and objectives are assumed to be efficient in terms of the overall (strategic) defence mission. Therefore, it is likely that, as a consequence of those analyses, the operational capabilities of our present military forces and the combat capabilities of many of today's sophisticated weapon

[1] Typically, in thqse studies the objectives are defined in terms of an array of targets which have to be neutralized within a specified time. As long as they survive, the weapon system are assumed to operate continuously at their technical and operational performance specification implying unlimited availability of targets, communication, POL and spares and assuming service support systems being attrited uniformly and at the same rates as the respective weapon system they support (see e.g. Huber and Neubecker [3]).

systems represent merely theoretical maxima. In reality, inadequate service support systems, less than perfect C^3I systems, and not quite appropriate tactics must be expected to degrade these capabilities perhaps significantly[2]. It is true that, in the past two decades, significant amounts of resources have been spent on improving the C^3I systems (Teates et al. estimate about 9 billion dollars have been spent up to 1980 in the US alone [1],p.40), but these efforts have not got anywhere near a situation that would justify the force designers' highly optimistic assumptions regarding the C^3I capabilities. On the contrary, some experts even believe that, by having considered C^3I in isolation and maximized the use of modern technology for its own sake, we may have ended up with systems that will not work in war (Cushman [5], p. 46).

A bureaucratic institutionalization of the action-oriented cost-effectiveness approach within the PPB management framework may have contributed, perhaps significantly, to such a situation brought about by piecemeal thinking in defence planning and analysis (see also [6,7]). In the planning phase of the PPB process, national goals are, through a top-down ends-means analysis, decomposed into elementary military tasks. For each of these tasks, alternatives to their accomplishment are determined in the programming phase. Based on a cost-effectiveness assessment of these alternatives the responsible planners compile program proposals for eventual implementation in the budgeting cycle.

However, due to the impact of modern technology military defence systems tend to become ever more complex; i.e., their elements are increasingly interrelated. They also represent open systems with a highly dynamic environment. Therefore, the decomposition of overall goals into subgoals and elementary tasks will hardly lead to set of truly disjunctive elements so that the "optimum" (most cost-effective) solutions determined separately for each of these tasks may, at best, be considered only as initial solutions that need to be iteratively tested for their mutual and environmental compatability in the light of resource constraints and enemy reactions. However, since military planners usually

[2] It is for this reason that Teates et al. reject the view of C^3I systems being considered "Force Multipliers" capable to somehow increase the effectiveness or even the numbers of a fixed set of weaponry. Because of the inseparable relationship between the C^3I system and the warfighting assets, the former is perhaps more of a "Force Divider" or rather a "Force Degrader" (see [2],p.30)

operate in an organizational environment characterized by rather rigidly compartmentalized hierarchical bureaucracies, the feedback necessary for an iterative testing and adaption of solutions hardly ever exists. But, without it, the holistic idea that, being systems analysts, the authors of the PPB approach undoubtedly attempted to implement in public planning, becomes perverted through the bureaucratic institutionalization of piecemeal thinking on the level of elementary tasks (i.e., the whole is the sum of its parts!). It also contributes to the planner's preoccupation with the future because the isolated problem solving makes him forget that his solutions may only be implemented in an evolutionary manner, i.e., they must be compatible with the present military force and its forseeable overall evolution.

As a consequence, force planning in most NATO nations seems to have largely degenerated into adjusting program proposals (independently arrived at by the services and arms branches) to the available resources. Unless political considerations dictate otherwise, resource constraints are usually effected through cuts across all of the elementary tasks with a bias toward the more visible weapon systems rather than support systems or even C^3I.

There is no doubt in my mind that, by attempting to be responsive to the daily needs of the military planners, the systems analysis community has contributed to this situation. As early as 1968 Schlesinger has postulated that we must get away from *traditional systems analysis* "..in the sense of analysis to assist in a simple choice between several given alternatives for accomplishing a single objective or task" ([8],p.387). An analysis approach is needed that accounts for the complementarities among the military missions and tasks, permits trade-offs among the conceptual and physical elements of the military forces, and explicetly considers the opponent's presumed reactions to the operational and structural planning options.

That the approaches of what Schlesinger called *traditional systems analysis* are still very much alive in military planning is obvious from the rather wide use of multilinear functions for the assessment of the effectiveness or utility of systems. A typical example can be found in a 1980 publication on a study to define the requirements for a new combat aircraft for the German Air Force [7,9]. There, the GAF mission was defined in terms of the number P_j of j-type targets (j=1,...m) that have to be neutralized within a specified time period in a conflict. With x_i denoting the number of weapon systems of type i and α_{ij} the number of j-type targets that an i-type system can be expected to neutralize within the specified time, the mission was defined by m linear equations

$$P_j = \sum_i a_{ij}\, x_{ij} \quad \forall j \tag{1}$$

where x_{ij} is the number of i-type weapon systems allocated to j-type targets. Using an LP, the (maximum) mission capability of the existing systems $i = 1,\dots,k$ was determined from

$$P_c = \sum_{j=1}^{m} \sum_{i=1}^{k} \alpha_{ij} x_{ij} = \text{Max} \tag{2}$$

such that Eq. (1) and the conditions

$$\sum_{j=1}^{m} x_{ij} \leq x_i \quad \forall i, \tag{3}$$

were satisfied. It turned out that

$$P_{cmax} = \sum_{j=1}^{l} P_j < \sum_{j=1}^{m} P_j \tag{4}$$

i.e., the maximum potential of the existing force was only sufficient to cover the targets of types $j=1,\dots l(l < m)$ leaving a deficit of

$$P_D = \sum_{j=l+1}^{m} P_j \tag{5}$$

as the mission for which a new system had to be specified. From the available alternatives $i = k+1,\dots,n$ the one with the lowest life-cycle cost C_i

$$C_i = x_i c_i (x_i/x_1,\dots x_k) = \text{Min}\,, \quad x_i = \sum_{j=l+1}^{m} x_{ij} \tag{6}$$

satisfying Eq. (5) was selected as the candidate alternative for the definition of the requirements. The creative contribution of systems analysis was more or less restricted to the piecemeal determination, target by target and system by system, of the target neutralization capability α_{ij}.

Another example for the use of multilinear functions in weapon systems planning can be found in a recent publication on the model employed by the German Navy for the assessment of the new frigate 122 (see [10], p.200). But also analysis engaged in net-assessment uses such linear functions.
An interesting example for this is Lucas Fischer's study of 1976 on the conventional balance of forces in Central Europe [11]. He measures the balance

by the quotient of two multilinear functions each expressing capabilities which the Warsau Pact and NATO are able to deploy as a function of time after the pact started mobilization:

$$P^* = \frac{\alpha_1 x_1 + \ldots + \alpha_i x_i + \ldots \alpha_n x_n}{\beta_1 y_1 + \ldots + \beta_j y_j + \ldots \beta_m y_m} \tag{7}$$

where x_i denotes the Pact and y_i the NATO inputs (such as soldiers and weapon systems), α_i and β_j the marginal capabilities of the respective inputs.

3. THE NEED FOR DYNAMIC ANALYSES THROUGH COMBAT SIMULATION

In order to illustrate the type of analysis required in military planning, Huber [7,12] has interpreted military defence as a production process. In analogy to economic theory he proposed an extended production function

$$P^* = f(x_1, \ldots x_i, \ldots x_n, y_1, \ldots y_j, \ldots y_m), \tag{8}$$

in which P^* denotes the defence product in terms of a net capability of the opponents X and Y, x_i and y_j their respective inputs. In fact, Eq. (7) is but a special functional form of such an extended production function.
When each side X and Y has but one type of input (homogeneous force structure), equation (7) reduces to

$$P^* = \frac{\alpha x}{\beta y} \tag{9}$$

describing the victory conditions for Lanchester-type battles between X and Y. Lanchester [13] hypothesized that under certain conditions attrition in combat between two homogeneous forces is, for each side, proportional to the numerical strength of the enemy. Thus, we may write:

$$\left.\begin{aligned} \frac{dx}{dt} &= -by \\ \frac{dy}{dt} &= -ax \end{aligned}\right\} \tag{10}$$

with x and y denoting the instantaneous numbers of live units of sides X and Y, and a and b the so-called attrition-rate coefficients. Eliminating dt from Eq. (10) and solving for the initial conditions $x = x_o$ and $y = y_o$, we obtain

$$a(x_o^2 - x^2) = b(y_o^2 - y^2). \qquad (11)$$

If we consider, e.g., X the winner, if $y = 0$ and $x > 0$, we readily deduce from (11) that

$$\frac{x_o}{y_o} > \sqrt{\frac{b}{a}}$$

or

$$K = \frac{x_o \sqrt{a}}{y_o \sqrt{b}} > 1. \qquad (12)$$

With $\alpha = \sqrt{a}$ and $\beta = \sqrt{b}$ Eq.(9) and Eq.(12) are identical.

This result shows that the military planner's production function is essentially a solution to a mathematical model of battle or war. However, for actual planning purposes Eq. (10) is hardly a sufficiently realistic model. To this end, the model must be considerably enriched (e.g.,accounting for hetereogeneous forces augmented by combat support elements and under conditions of variable attrition-rate coefficients) in which case analytical solutions are rather impossible. The only feasible approach to eventually specify military production functions is the *experimental* one through battle simulations.

Attempts to derive such functions from records of historical conflicts and/or personal combat experience are usually frustrated by the randomness of combat circumstances and by insufficient records and recollections. But even if we had perfect records, the usefulness of production functions derived from them would be rather questionable.

This is because we can not be sure whether the historical processes represent *efficient* processes in the sense that both sides employed their forces in a mutually "optimal" manner. The implicit efficiency assumption of economic production theory (that the production processes from which the production function and its coefficients are determined) is only justified as long as a competing market eliminates inefficient producers. However such an *a priori* assumption is highly questionable in the defence field because (1) defence organizations meet few of the characteristics of competitive firms; (2) their ultimate objective is nothing less than trying to prevent the "market" from happening. Thus, an explicit analysis of the processes by which military inputs are converted into outputs, that is an examination and eventual adaption of the operational principles and doctrines controlling the military production processes, is an indispensable task within military force planning. Technological change makes this an even more urgent requirement, even if combat experience is available. Christopher Harvie's essay on "Technological Change and Military Power in Historical Perspective" [14] presents ample historical evidence on both, how the adaption of operational principles to new technology provided the decisive edge, and how a retrospective military ideology stood in the way of innovation[3].

Steven Canby's 1973 criticism of the systems analysts around Alain Enthoven mainly concerns the efficiency assumption implied in their analyses. They accepted the philosophies, concepts, and operational principles underlying military organizations as *given* rather than making them subject of the analysis. That Enthoven considered them to be efficient is obvious from his argument: "Army force planners must be satisfied with the current force structure because they have not proposed changes when invited to do so" (see [4],p.9).

In the context of long range armaments planning Huber [15] discusses an example which demonstrates the impact of operational concepts on capabilities and structures of two opposing tactical air forces. Based on the assumption that tactical air operations are performed in direct and/or indirect support of the land battle, he considers tactical air war as a multistage game where the adversaries decide, at each stage, how to allocate their tactical aircraft to the basic tactical air missions of Offensive Counter Air (OCA) and Offensive Air Support (OAS), so that their respective capabilities to support the land battle become a maximum relative to that of the enemy. With respect to force structure planning, both of the two fictitious antagonists are considered to be constrained by constant budgets.

[3] The German *Blitzkrieg* strategy of World War II, combining a mass tank offensive with infantry and air support, is the classical example for both. It had been anticipated by Fuller and Lidell Hart, but it was dismissed by the traditionalists in the British Army. But in Weimar Germany, the hard restrictions of the treaty of Versailles enforced a drastic break with the past. It encouraged the development of new options which became evident in the early campaigns of WW II. However, the operational and technical innovations of the Germans in the land/air war domain were not matched in naval warfare. In fact, their naval construction plans were quite old-fashioned, an indication that the navy's operational thinking essentially resumed where it had ceased after World War I.

Tab. 1 shows results of four games in terms of the relative OAS-capabilities of defender (V) versus attacker (A)[4]. The underlying force structures are characterized by heavy (H) and medium (M) tactical aircraft on which both sides spend half of their available budgest each. A is assumed to have 30% more aircraft, V to have aircraft of somewhat better performance with respect to payload and weapon effects as well as sortie capability.

		attacker A	
		no OCA	opt. OCA
defender V	no OCA	0.86	0.99
	opt.OCA	2.87	1.0

Tab. 1: Relative OAS-capabilities of V versus A

The values in Tab. 1 are normalized around the case where both sides pursue their optimum policies in terms of mission allocations (opt. OCA). Thus, all values must be interpreted in relation to the "balance" of that case. Whatever actual OAS-capability ratio may have resulted there, the values indicate that it reduces to 86% if both sides allocate their aircraft to OAS only. The values also show that the attacker A is practically forced to open the campaign with an OCA operation. If A leaves the OCA to V, then the OAS-potential ratio increases by a significant factor (2.87). This result is due to the superior OCA performance characteristics of V's systems which A must prevent from becoming effective. This is true in spite of the low OCA effectivenss of A's systems which do not significantly reduce the OAS ratio (from 1 to 0.99) when V does not react to A's OCA.

[4] The terms *attacker* and *defender* do not imply that the respective antagonist is limited to offensive or defensive operations. They only indicate that the attacker initiates the hostilities, i.e., gets the first move, to which the defender reacts. Both have perfect intelligence.

This example illustrates that, for given force structures on both sides, the balance may vary significantly depending on the operational concepts chosen by each side. This is to say that the efficiency of military production processes is not only depending on one's own operational principles and doctrines, it is equally sensitive to those of the potential enemy. Quite similar findings were recently presented by Farrell [16] who has evaluated a large number of simulated land combat histories. He showed that, in many instances, tactics and doctrine have a much more decisive influence on combat outcome than weapon system performance parameters.

The implications of these findings for the systems analysis supporting the planning of military systems are obvious: Military systems and their elements may be adequately assessed only through dynamic analyses employing, in an experimental fashion, gaming models[5], which "act out" combat and provide the information necessary to trade-off resources (men,systems), structure,doctrine and tactics explicitly accounting for the same factors on the potential opponent's side. Being the art of employing combat resources, tactics and doctrine are implemented through command decisions requiring some kind of C^3I capability. Therefore, a more or less explicit representation of C^3I in combat simulations is prerequisite to (1) a more realistic assessment of the capabilities of the existing forces; (2) the evolution of tactics and doctrine such that the inherent force capabilities can be fully exploited; and (3) the "balanced" design of military systems in general and of C^3I system in particular.

[5] The Term "gaming" is used to characterize two-sided battle models in which the opponents react to each others actions either through interaction of human commanders or through a formalized contingency logic or through simple decision rules. Such games comprise the entire range from *military exercises* and *combat experiments* employing men and equipment to the highly abstract *analytic games*. The above discussed air war game is an example of an analytic game using simple decision rules. It describes the states x^μ and y^μ of the opposing air forces at time step μ by two state equations

$$x^{(\mu+1)} = f(x^{(\mu)}, y^{(\mu)}, \alpha_\mu,\beta_\mu)$$

$$y^{(\mu+1)} = f(x^{(\mu)}, y^{(\mu)}, \alpha_\mu,\beta_\mu)$$

where α_μ and β_μ denote the operational strategies of the adversaries. The game theoretic decision rule is given by

$$\min_{\alpha_\mu \in E_A} \max_{\beta_\mu \in E_D} U_\mu$$

where $U_\mu = I_D(\alpha_\mu,\beta_\mu)/I_A(\alpha_\mu\ \beta_\mu)$ is the utility function with I_D and I_A the offensive air support potentials of defender and attacker respectively.

4. INTERACTIVE SIMULATION FOR C^3I RESEARCH

It seems to be generally true that the acceptance of models of socio-technical systems tends to decrease as the degree to which human factors influence the system dynamics increases (see,e.g.,Schultz and Slevin [17]). The evidence available from military OR/SA certainly attests to that. With air combat being influenced to a much larger extent by physical and engineering than by human processes, at least when compared to land combat, air war models have become accepted earlier and to a much higher degree than land war models[6].

But a similar pattern can be recognized within land combat modelling if we ackowledge that the acceptance of models is closely correlated to their state-of-the art. Of the six combat processes usually distinguished in the literature (see,e.g.,Low [18]), modelling of *attrition* and, to a somewhat lesser degree, of *movement* is much more advanced than modelling of *suppression, combat support, combat service support,* and of C^3I (see Huber [19]). With regard to the latter, last year's NATO-symposium on "Modelling and Analysis of Defence Processes" concluded, that ".. for those processes that are well understood in the sense of physics and engineering, there are quite adequate models available. These include communications, the electronic effects of ECM and decoys, collection system performance, and computer processing. The weakest link in modelling C^3I processes and systems is the human element. Not much is known about the higher order cognitive functions and the population of decision makers who implement tactics and doctrine and respond to intelligence and ECM" ([19],p.15).

In a historical perspective, this situation seems not surprising. Because it was the very ignorance of how tactical decision makers operate in a more complex combat environment that made the systems analysts discover the rather old military tool of interactive gaming[7]. By the incorporation of a human

[6] The history of military OR/SA in Germany is proof of that. It started in 1962 when air force and navy initiated the operations of analysis groups to be followed by the army only about half a decade later. Also, while air force and navy emphasized combat modelling right from the start, most of the initial army studies were related to logistics.

[7] In most closed combat simulation models and in the analytic games the combat environment is rather simple. A typical example is the above discussed analytical air war game. There, the opposing commanders only decide on the proportion of their forces allocated to Offensive Counter Air and to Offensive Air Support. The model assumes that, in doing so, they have perfect information on the instantaneous states and the histories of their own and the enemy's resources. Also, practically the entire C^3I system is represented by the min-max decision rule.

gamer to represent the tactical decision maker, the problem of inadequate decision models was circumvented. But,by doing this, C^3I became an implicit element of the military system models that did not readily lend itself to being traded-off versus other elements. In interactive combat simulations gamers and game controllers more or less "simulate" the C^3I systems.

The development of formal models of C^3I systems was part of more or less isolated assessments that were mostly restricted to the two lower system levels defined by Alberts [20], that of *technical system performance*[8] and that of *information attributes*[9] (see Fig. 2). If performed at all, assessments on the third level of *information value* usually assumed, in true piecemeal fashion, that the information value increases monotonically as the information attributes improve (motto: more and faster is better than less and slower).

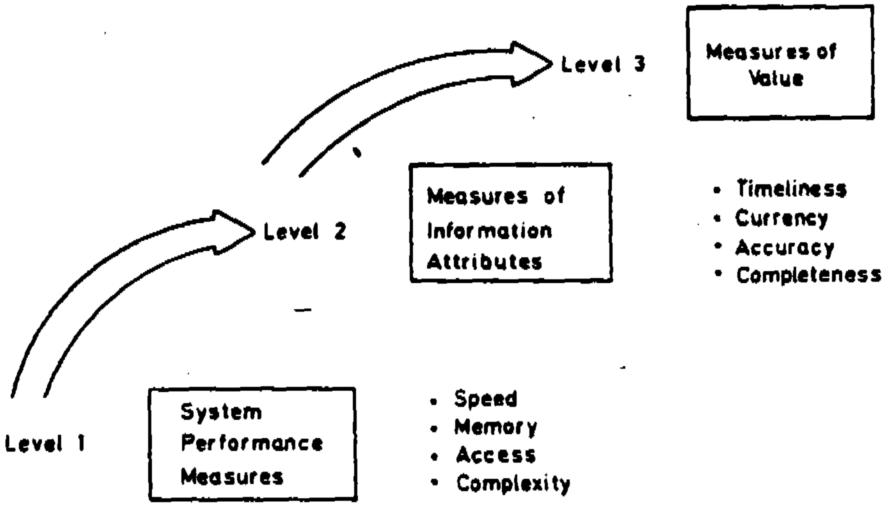


Fig. 2: Task Specific Measurement Levels of C^3I Systems (Alberts 1980)

There have been several proposals on how to measure the information value of C^3I systems (see,e.g.,Cushman [5], Alberts [20], Miller [21], Huber and

[8] E.g., communications speed, memory size, access time, instruction complexity, and I/O characteristics.

[9] Timeliness, currency, accuracy, completeness, and ease of use.

Hofmann [22]). They all agree that such measures must somehow express the consequences as to the expected course and outcome of the operations which the respective C^3I systems are serving. In the context of battlefield C^3I, this would be best accomplished through battle simulation [10].

But since data become information only after being processed in the recipient's mind, such simulation experiments need to be *interactive*, i.e., they must include the tactical decision maker. Interactive simulation would also permit to make the cognitive processes of military commanders object of exploratory research thus helping to close the above indicated gap of knowledge on the human element in C^3I systems. And last but not least, interactive simulation is a *conditio sine qua non* for the evolutionary development and the test of decision support systems and the application of artificial intelligence in battlefield C^3I.

Indeed there is empirical evidence that seems to underline the desirability of some decision support capability for the field commander. From series of 69 interactive combat games involving 23 sets of players Daniel [23] arrives, among others, at the conclusions that (1) more data does lead to better quality decisions, though the effect is small compared to the variations in results between different players; (2) prior intelligence (as opposed to intelligence obtained throughout the simulated battles) obviously dominates decisions; (3) players who make the "best" decisions take considerably longer than average to play the game (in fact, Daniel's slowest player took twice as long as the fastest player); and (4) the "poorer" players do lot better with more data then the "better" players do with less data.
According to Daniel, the question as to whether the rather small impact that data levels (of in-battle intelligence) had on the quality of player decisions is symptomatic of player's inability to make use of the data, or merely reflects the fact that high data levels are perhaps superfluous, is yet to be answered. But either way, we might conclude that merely providing more and more current data may, for battlefield C^3I, yield only disappointing returns. Rather, commanders should be given some data "processing" capability thus providing them effectively more time for their decisions.

This suggests that, contrary to the hitherto practiced philosophy

[10] We concur with Cushman's observation that ".. a program of battle simulation will also foster incremental, evolutionary growth of C^3 systems, through a process of systematic trial and modification, in the absence of conflict ([5],p.47).

of designing highly centralized C^3I systems in the fashion of the classical management information systems, we perhaps ought to pursue a highly decentralized architectural approach with some AI capability. Such a philosophy would also permit pursuit of the evolutionary growth of C^3I systems postulated by Cushman because the necessary battle simulations could be largely performed as part of routine training exercises using the processors of such a decentralized C^3I system.

Indeed, the so-called command and staff simulators (CSS) proposed by Huber [7,12,24] as part of a comprehensive systems analysis approach in support of force planning[11] could be gradually materialized as part of such a C^3I evolution. The CCS would be basically designed as interactive computer games providing military staffs, at all command levels, a dynamic (combat) environment for their work. "In addition..., CSS-systems would also permit to better assess staff performance. They would provide continuously updated information on command and control cycles as well as a readily available testbed for command and control systems" (see [12],p. 107).

As of today, we[12] have through a series of theses, developed the basic software package for a battalion/brigade-level CSS to demonstrate the feasibility of such systems in form of a portable prototype simulator to be developed within a research program (hopefully) supported by the German Army. In addition to providing a training tool, this simulator is conceived as an instrument for empirical research on tactical decision processes with a view to the development of decision support systems in C^3I (see [25],p.4).

5. TWO EXAMPLES

In order to illustrate the type of problems to be tackled through gaming experiments by means of interactive combat simulation, two recently

[11] I.e., the "Compound Gaming Approach" employing hierarchically ordered families of interacting combat models of both, the formal and the physical kind, which permit addressing future issues and alternatives in the light of current capabilities and deficiencies.

[12] Together with Prof. Hans W. Hofmann at the Institute of Applied Systems and Operations Research of the German Armed Forces University.

published examples shall be briefly reviewed. One is on the use of gaming to develop *cognitive maps* as means to establish the way in which tactical commanders model their (subjective) decision environment. The other describes an approach to deriving, based on information obtained from interactive gaming experiments, a *decision model* for tactical situations in which a multitude of criteria have to be considered.

5.1 Cognitive Mapping

In the 1982 paper on "Cognitive Maps of Decision-Makers in a Complex Game" [26], Klein and Cooper report on a series of manual gaming experiments in which a number of players acting as divisional commanders were confronted with two scenarios each, a defence scenario and a advance-to-contact scenario. The players believed to be part of a team (consisting of themselves and of one superior and two subordinate commanders) playing interactively against a purposeful enemy. But,without the divisional commanders knowing this, the enemy and divisional players' team-mates were played by the game controllers with their actions entirely predetermined. Thus, the players could be led through the same sequence of pre-planned events in the game, so that their behaviour could be compared under an identical sequence of objective circumstances. During the course of the game, in each time period, the players had to make reports to their superior commanders and to issue directions and orders to their subordinate commanders. This communication was taped and from the transcripts of the recordings, cognitive maps were derived for each player .

"A cognitive map is a representation of the perceptions and beliefs of an individual about his own subjective world" ([26],p.63). It depicts the *concepts* used by the individual and the *causal relationships* between them. In their experiments, Klein and Cooper only considered two types of relationships, positive and negative. A *positive* relationship exists when a change in the predecessor concept causes a similar change in the successor. A *negative* relationship characteristics a case where an increase (decrease) in the predecessor causes a decrease (increase) in the successor. As an example, the player statement that "..morale is high, as they (members of a friendly unit) are advancing with little opposition..." is analyzed to exhibit a positive relationship between the concepts of *"high morale"* and *"unopposed advance"*. As an example, Fig. 3 shows the cognitive maps thus derived for too different players in the advance-to-contact scenario.

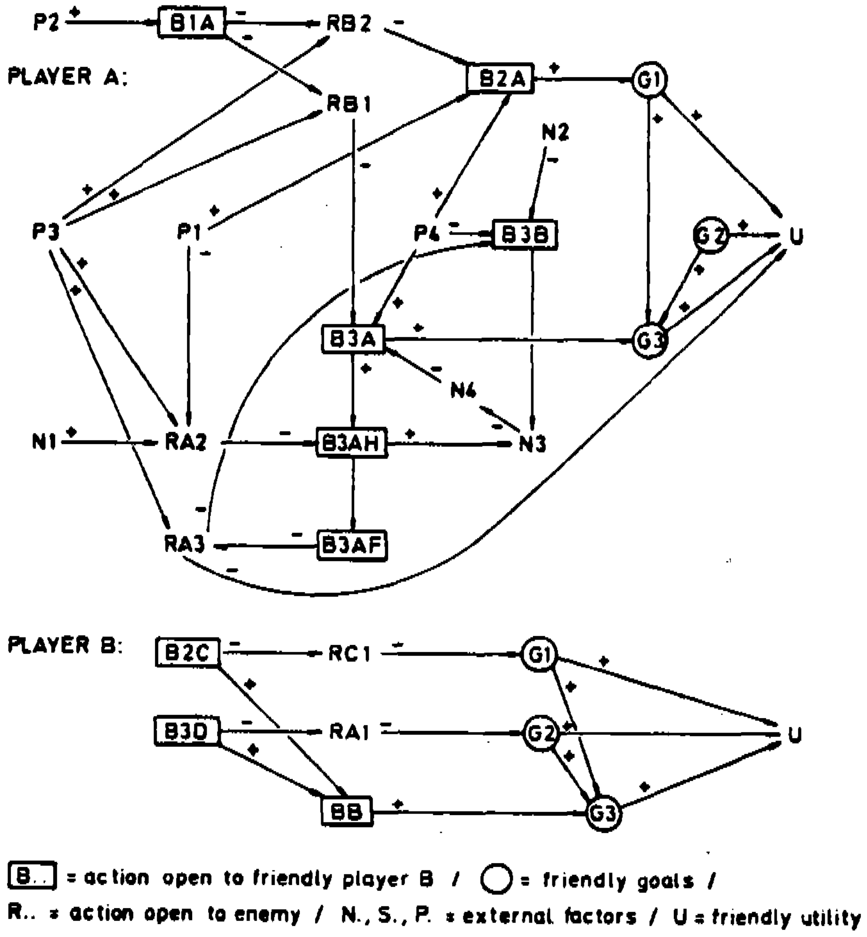

P2
B1A
RB2
PLAYER A:
B2A
G1
RB1
N2
P3
P1
P4
B3B
G2
U
B3A
G3
N4
N1
RA2
B3AH
N3
RA3
B3AF
PLAYER B:
B2C
RC1
G1
B3D
RA1
G2
U
BB
G3
B.. = action open to friendly player B / ○ = friendly goals /
R.. = action open to enemy / N., S., P. = external factors / U = friendly utility

Fig. 3: Cognitive Maps in Advance-to-Contact Scenario
(Klein and Cooper [27])

From such maps Klein and Cooper noticed, among others, a rather significant difference in the number of concepts identified by different players and in the densities[13] of their cognitive maps. But, for the majority of players, the number of concepts was quite similar in the two scenarios. This leads them to conclude that the number of concepts ".. has apparently little to do with the objective situation and may present some limit to the quantity of concepts that the decision-maker feels he can usefully cope with at any one time" [26],p.66).

[13] Number of observed links divided by the maximum number of possible links.

But the fact that the larger maps generally exhibit a comparatively low density seems to indicate that larger maps contain more peripheral concepts of limited influence. "..the central sections of different players' maps are ... appearing in the same or slightly altered form in several players' maps." ([26],p.66).

From these and other results it appears that cognitive mapping should be a valuable tool for the structuring of *knowledge bases* in tactical expert systems and for organizing their *data bases* through a series of properly designed combat experiments. Using interactive computer-simulation, it should also become possible to shed some light on the largely unresolved issue of decision quality as a function of map size and density and on the impact of training and doctrine with regard the adequacy of decisions in given scenarios.

5.2 Decision Modelling

The idea of using interactive combat simulation to develop descriptive decision models for incorporation into closed combat games has been proposed by Reidelhuber [28,29]. But it could also be employed for the development of decision aids in tactical C^2 systems.

Reidelhuber interprets the tāctical decision problem as having to choose the action A_D which is the (most) appropriate in a given decision situation described by a data vector

$$X_D = \{x_{D1},\ldots,x_{Dj},\ldots,x_{Dm}\} \tag{13}$$

of m criteria that may be thought of as a two dimensional profile as shown in Fig. 4a.

In order to find A_D, Reidelhuber assumes that, in a specific decision situation, there is a known set of possible actions $A_i(i=1,\ldots,n)$ and a set of data vectors

$$\bar{X}_i = \{\bar{x}_{i1},\ldots,\bar{x}_{ij},\ldots,\bar{x}_{im}\} \quad (i=1,\ldots,n) \tag{14}$$

each of which is *reference* profile for the corresponding A_i (see Fig. 4b). The decision problem thus reduces to the question as to which A_i is best in situation X_D and should be selected as A_D.

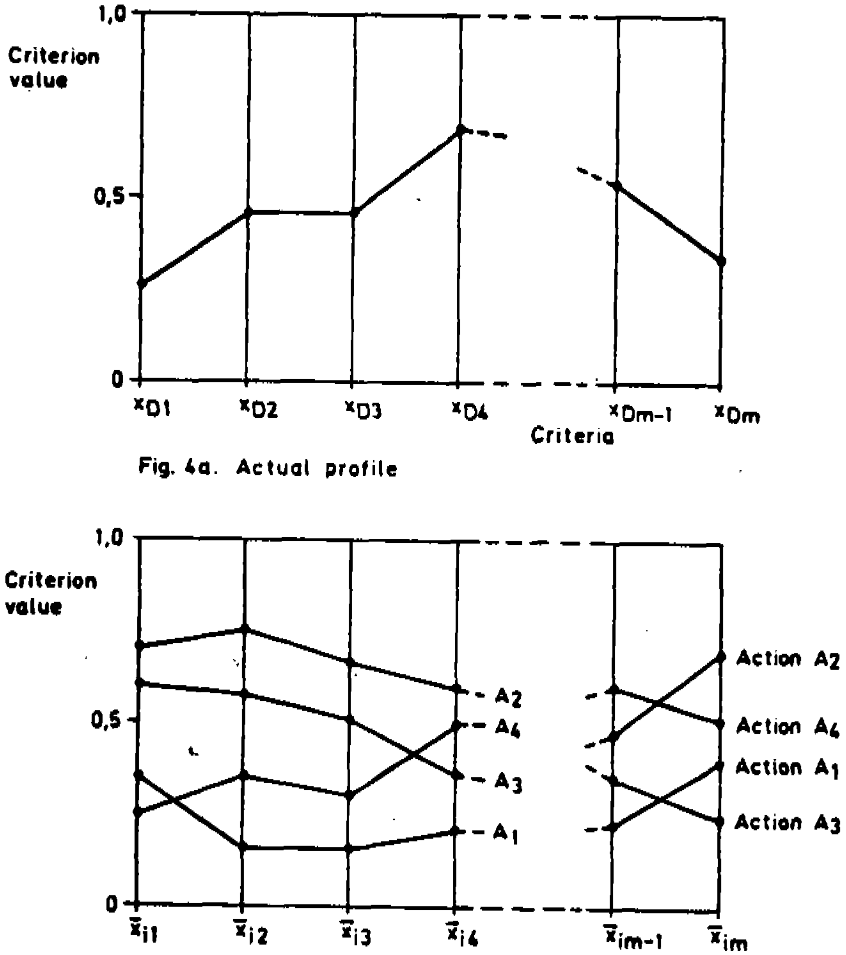


Fig. 4a. Actual profile

Fig. 4b. Reference profiles

Fig. 4: Profiles in a Decision Situation (Reidelhuber 1982)

To this end, Reidelhuber proposes the decision rule

$$\min_{i \in I} P(A_i) = \sum_{j=1}^{m} r_{ij}^2 (\bar{x}_{ij} - x_{Dj})^2 \rightarrow A_i, \qquad (15)$$

where r_{ij} denotes the relevance factor taking into account that different criteria may have different weights in the decision considerations.

For the determination of the reference profiles $\bar{X}_i$ and the relevance factors r_{ij} a number of p similar decision situations (p >> n) is generated in course of an interactive combat simulation to find p data vectors thus establishing a matrix X of primary data.

$$X = \begin{bmatrix} x_{11},\ldots,x_{1j},\ldots,x_{1m} \\ , \\ . \\ x_{k1},\ldots,x_{kj},\ldots,x_{km} \\ . \\ , \\ x_{p1},\ldots,x_{pj},\ldots,x_{pm} \end{bmatrix} \qquad (16)$$

Furthermore, of the n actions A_i taken by the players, one is assigned as the best to each of the p situations, i.e., to each row in the matrix X. Then all data vectors with the same action assigned are collected in disjunctive classes. For each action A_i, the reference profile $\bar{X}_i$ is determined as the vector of the mean criteria values $\bar{x}$ of the class members

$$\bar{X}_i = \frac{1}{q_i} \{ \sum_{l=1}^{q_i} x_{l1}(A_i),\ldots, \sum_{l=1}^{q_i} x_{lj}(A_i),\ldots, \sum_{l=1}^{q_i} x_{lm}(A_i) \} \qquad (17)$$

with q_i as the number of members assigned to class i.

The relevance factor r_{ij} is defined as the reverse of the scaled standard deviation s^*_{ij} of the criteria values found within each class i.e.

$$r_{ij} = \frac{1}{s^*_{ij}} , \qquad (18)$$

$$S^*_{ij} = \frac{s_{ij} + \frac{1}{m} \sum_{l=1}^{m} s_{il}}{\sum_{g=1}^{m} (s_{ig} + \frac{1}{m} \sum_{l=1}^{m} s_{il})} = \frac{s_{ij} + \bar{s}_i}{2m\,\bar{s}_i} , \qquad (19)$$

$$s_{ij} = [\frac{1}{q_i - 1} \sum_{l=1}^{q_i} (\bar{x}_{ij} - x_{lj}(A_i))^2]^{1/2} , \qquad (20)$$

with at least one $s \neq o$ in each class. The idea of this definition is that the relevance of a criterion for the decision maker is the smaller the more strongly the criterion value deviates when he selects the same action. The transformation means normalizing s and shifting the zero point, because no relevant criterion should have a weight that is too small or too large. The value region of s^*_{ij} is

$$\frac{1}{2m} \leq s^*_{ij} \leq \frac{m+1}{2m} .$$

Fig. 5 shows Reidelhuber's concept for the development and adaption of the decision model to new weapon systems, to alternative engagement tactics, and to modified command and control doctrines. It starts with the existing decision structure and ends with the test of the decision model in a "closed" simulation (decision model replaces player). In case the results are not acceptable, the player may interactively modify the decision model.

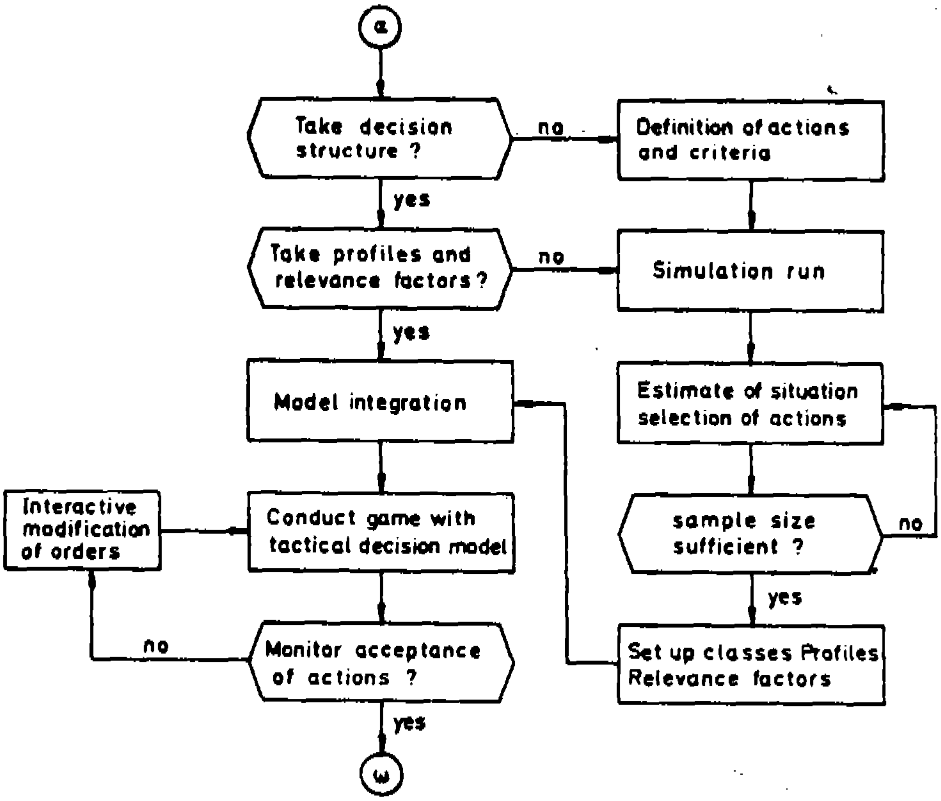

a
Take decision structure ?
no
Definition of actions and criteria
yes
Take profiles and relevance factors ?
no
Simulation run
yes
Model integration
Estimate of situation selection of actions
Interactive modification of orders
Conduct game with tactical decision model
sample size sufficient ?
no
yes
no
Monitor acceptance of actions ?
Set up classes Profiles Relevance factors
yes
ω

Fig. 5: Concept for the Development of the Tactical Decision Model
(Reidelhuber 1982)

6. FINAL REMARKS

In this paper, the attempt was made to 1) demonstrate the necessity for the employment of interactive combat simulation in order to provide the information for a truly holistic assessment of military systems in general and C^3I systems in particular; 2) illustrate how interactive gaming could assist in developing and testing models of cognitive processes in C^3I, thus, not only helping to close a fundamental modelling gap, but also aiding the evolution of new systems with some AI capability.

To this end, the gradual implementation of a "compound gaming approach" is suggested, because in the absence of military conflicts there is hardly a viable alternative to assure an *adaptive control* of the development of military forces and systems in a mission oriented context and with due regard to the limited availability of resources. In particular with a view to a *"balanced"* development in the sense of a *robust* combination of weapon systems mix, force structure, tactics and operational concepts, means must be available to study all of the force elements in an interdependent fashion so as to show the possibilities and the limits of their mutual substitutability. There is some indication that piecemeal thinking and analysis has led us to opt rather strongly for technology resulting in systems that are perhaps not very robust and rather expensive so that we may soon no longer be able to afford force sizes that make sense operationally.

ACKNOWLEDGEMENTS

The preparation of this paper was supported by the U.S. Army Research, Development and Standardization Group (UK) under contract No. ERO-70-83.

References

[1] Teates,H.B. et al.: Defining and Measuring C^4-C^3I-C^3-C^2I-C^2, Part One: A Perspective. Military Electronics/ Countermeasures, May 1980,pp.40-46,92-93

[2] Teates,H.B. et al.: Defining and Measuring C^4-C^3I-C^3-C^2I-C^2. Part Two: System Evaluation. Military Electronics/Countermeasures,June 1980, pp. 28-34

[3] Huber, R.K., Neubecker,K.A.: Operations Research für die Luftwaffe-Bewertung von Flugzeugwaffensystemen. Jahrbuch der Luftwaffe Folge 10,1973,pp.68-74

[4] Canby,S.L.: NATO Military Policy: Obtaining Conventional Comparability with the Warsaw Pact. Rand Report R-1088-ARPA, June 1973

[5] Cushman,J.H.: Exercise, Field Test and Experimentation and Battle Simulation Approaches. Proceedings of the second Conference/Workshop on the Quantitative Assessment of the Utility of C^3 Systems, October 1979, pp.28-34

[6] Cover·Story: The Winds of Reform. Time Magazine, March 7,1983, pp.9-18

[7] Huber,R.K.: Die Systemanalyse in der Verteidigungsplanung - Eine Kritik und in Vorschlag aus systemanalytischer Sicht. Wehrwissenschaftliche Rundschau Heft 50,1980,pp.133-143

[8] Schlesinger,J.R.: The Changing Environment for Systems Analysis. In: Systems Analysis and Policy Planning - Applications in Defence (Quade, Boucher Eds.) Elsevier, New York 1975,pp.364-387

[9] Flume,W.: TKF-noch nicht entscheidungsreif-aber Industrie ist vorangekommen. Wehrtechnik 5(1980),pp.30-33

[10] Molitor,H.P.: Bewertungsmodell Fregatte 122-Pilotprojekt für Großwaffensysteme. Soldat und Technik 4/1982,pp.200,207

[11] Fischer,L.: Defending the Central Front: The Balance of Forces. Adelphi Paper No. 127, The International Institute for Strategic Studies, London 1976

[12] Huber,R.K.: A Systems Analyst's View on Force Structure Planning. Lectures on Systems Analysis, KIDA 1980, pp.67-116

[13] Lanchester,F.W.: Aircraft in Warfare: The Dawn of the Fourth Arm - No.V, The Principle of Concentration. Engineering 98(1914),pp.422-423

[14] Harvie,C.: Technological Change and Military Power in Historical Perspective. pp. 5-13 in: New Conventional Weapons and East-West Security, Part I, Adelphi Paper No. 144, The International Institute for Strategic Studies, London 1978

[15] Huber,R.K.: Ein analytischer Ansatz zur Untersuchung langfristiger Zielvorstellungen der Luftrüstungsplanung. In: Operationsanalysen für die Luftrüstung (Huber,Dathe Hrsg.) München-Wien 1978,pp.161-178

[16] Farrell,R.L.: How Non-Weapon System Parameters Affect Combat Results. Proceedings of the 1982 NATO-Symposium on Modelling and Analysis of Defence Processes. To appear

[17] Schultz,R.L. Slevin,D.P.(Eds.): Implementing Operations Research Mangement Science. New York-London-Amsterdam 1975

[18] Low,L.J.: Theater-Level Gaming and Analysis Workshop for Force Planning. Vol.II - Summary, Discussion of Issues and Requirements for Research.SRI-Report, May 1981

[19] Huber,R.K.: Synthesis of the Findings of the Symposium on Modelling and Analysis of Defence Processes. Proceedings of the 1982 DRG-Seminar on operational Research for the Selection and Design of Future Military Systems. To appear

[20] Alberts,D.S.: C^2 Assessment: A Proposed Methodology. Mitre Corporation. Jan 1980

[21] Miller,R.S.: On the Assessment of Command and Control. Führungs- und Informationssysteme, München-Wien 1982,pp.324-368

[22] Huber,R.K., Hofmann,H.W.: Zum Problem der Bewertung von Einsatzführungssystemen. Führungs- und Informationssysteme, München-Wien 1982, pp.299-323

[23] Daniel,D.W.: What Influences a Battlefield Command Decision. Proceedings of the 1982 NATO-Symposium on Modelling and Analysis of Defence Processes. To appear

[24] Huber,R.K.: The Systems Approach to European Defence - A Challenge for Operational Research Gaming. Phalanx, Sep 1982, pp.18-21

[25] Hofmann, H.W., Huber R.K.: Antrag auf Forschungsförderung für das Projekt "Entwicklung eines Gefechtssimulators". Neubiberg, March 1983

[26] Klein,J.H., Cooper,D.F.: Cognitive Maps of Decision-Makers in a Complex Game. Operational Research, Vol. 33 No.1, Jan.1982,pp.63-71

[27] Cooper,D.F. et al.: The Development of a Research Game. Operational Research Vol. 31 No. 2, Feb. 1980 , pp. 191-193

[28] Reidelhuber,O.: Ansätze zur Entwicklung eines Entscheidungsmodells für Gefechtssimulationen. Operationsanalytische Spiele für die Verteidigung, München-Wien 1979,pp. 257-281

[29] Reidelhuber,O.: Modelling of Tactical Decision Processes for Division-Level Combat Simulations. Proceedings of the 1982 NATO-Symposium on Modelling and Analysis of Defence Processes. To appear

DECISION ALGORITHMS IN FUZZY SITUATIONS

H.-J. Zimmermann, Institute of Technology
Templergraben 55, 5100 Aachen (FRG)

ABSTRACT. Classical decision models and algorithms are either dichotomous in character (feasable - nonfeasable, optimal - not optimal) or they are stochastic. There are, however, many decision making situations which are ill-structured or vague and which cannot properly be modelled by use of classical tools. Fuzzy set theory has been put forward as a possible bridge between models and reality in above mentioned vaguely described situations. In the meantime more than 4000 publications are available in the area of fuzzy set theory and its applications. Of particular interest for fuzzy set modelling seems to be the area of decision making. Algorithmic approaches such as fuzzy linear programming as well as results of axiomatic and empirical research and their application to civilian and military problems are represented.

1. INTRODUCTION Most of our available models and algorithms are "crisp", i.e. based on traditional mathematics or dual logic which are both dichotomous in character. This is certainly appropriate if the problem under consideration is of the yes-or-no type as frequently encountered in the physical sciences, in engineering or in hardware design. Here we can clearly distinguish between optimal and nonoptimal, feasable and nonfeasable solutions etc.

Decision making normally involves human judgments, evaluations and perceptions. Their structure is not dichotomous but rather vague, not of the yes-no but of the more-or-less type. If an essentially two-values modelling language is used to model this type of problem then in-appropriate models may result and model-solutions do not coincide with problem solutions.
There are different types of vagueness which have to be taken into account:

I. Vagueness concerning the occurrence of crisply described events. This is and has been the domain of probability theory. Typical statements of that type are: "The probability of hitting the target is .6" or "There is a good chance of meeting him".

II. Vague Phenomina

a.) Intrinsic Vagueness

This is a type of vagueness which is due to the vagueness of human judgments and concepts. Examples are terms such as "tall men", "acceptable profits", "high vulnerability", "long sticks" etc.

b.) Informational Vagueness

We are accustomed to the view that a lack of information causes vagueness (for instance in stochastic statements). There is, however, also a type of vagueness which is due to an abundance of information. A "creditworthy" person could, for instance, be fully described by using a large number of "descriptors". Since the human capacity for information processing and storage is very limited, not all the necessary descriptors will be in the mind of a person when using the term creditworthiness. Human beings can still communicate using these terms which are generally called "subjective categories". But when doing so the set of "creditworthy men" is no longer a set in the classical sense but rather a category with vague bounderies. Decomposing such terms normally makes the subcategories better and sharper defined because fewer descriptors are needed to describe them sufficiently and men can be aware of a larger fraction of these descriptors at the time when using the terms. The next three pictures despict the hierarchies of "Subjective categories" and their subcategories.

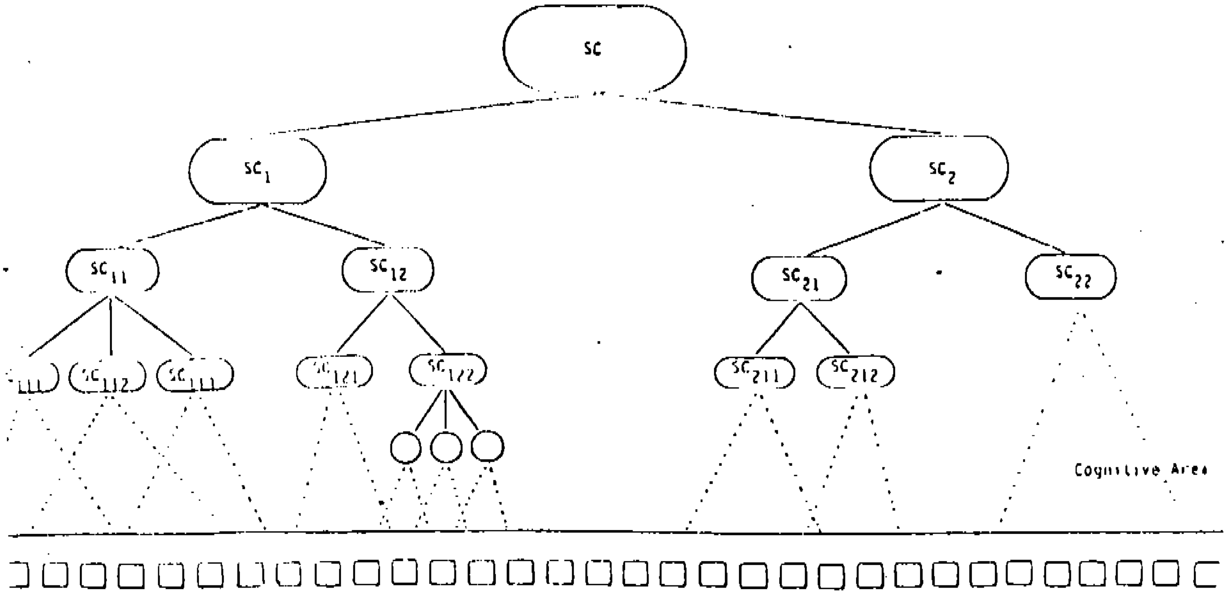


Figure 1: Evaluation hierarchy (SC = Subjective Category)

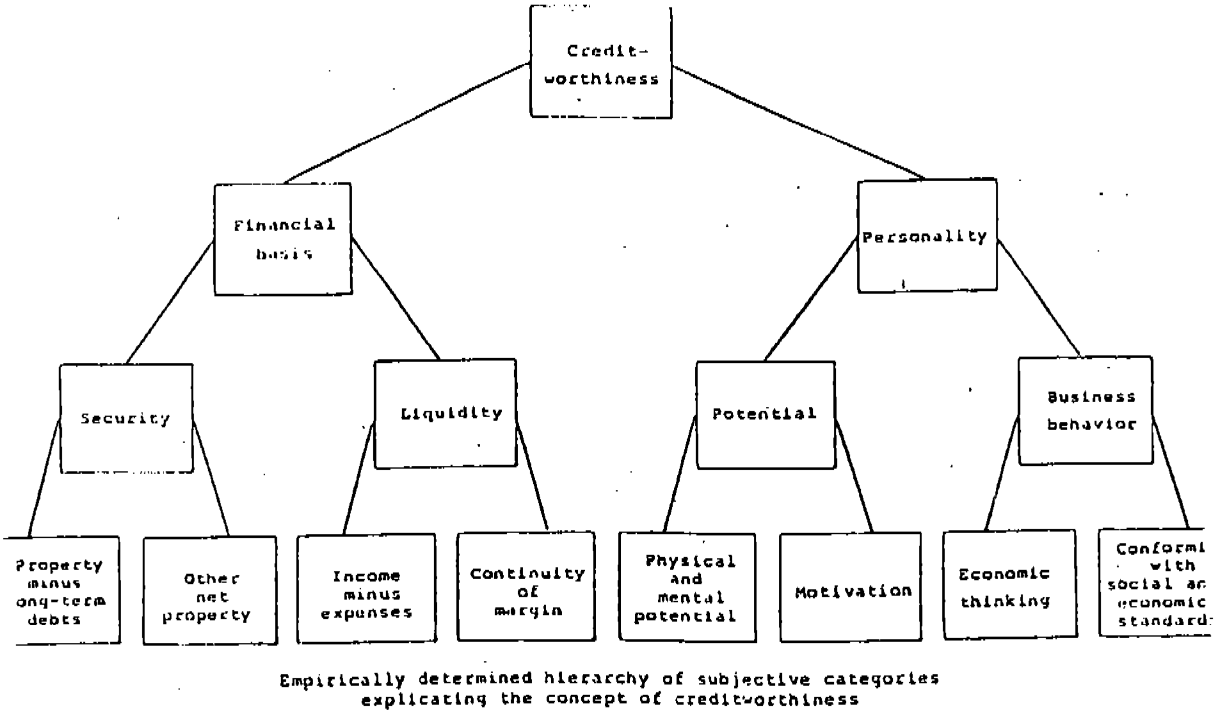

Credit-worthiness
Financial basis
Personality
Security
Liquidity
Potential
Business behavior
Property minus ong-term debts
Other net property
Income minus expenses
Continuity of margin
Physical and mental potential
Motivation
Economic thinking
Conformi with social ar economic standard
Empirically determined hierarchy of subjective categories explicating the concept of creditworthiness

Figure 2: Empirically determined hierarchy of subjective categories explicating the concept of creditworthiness

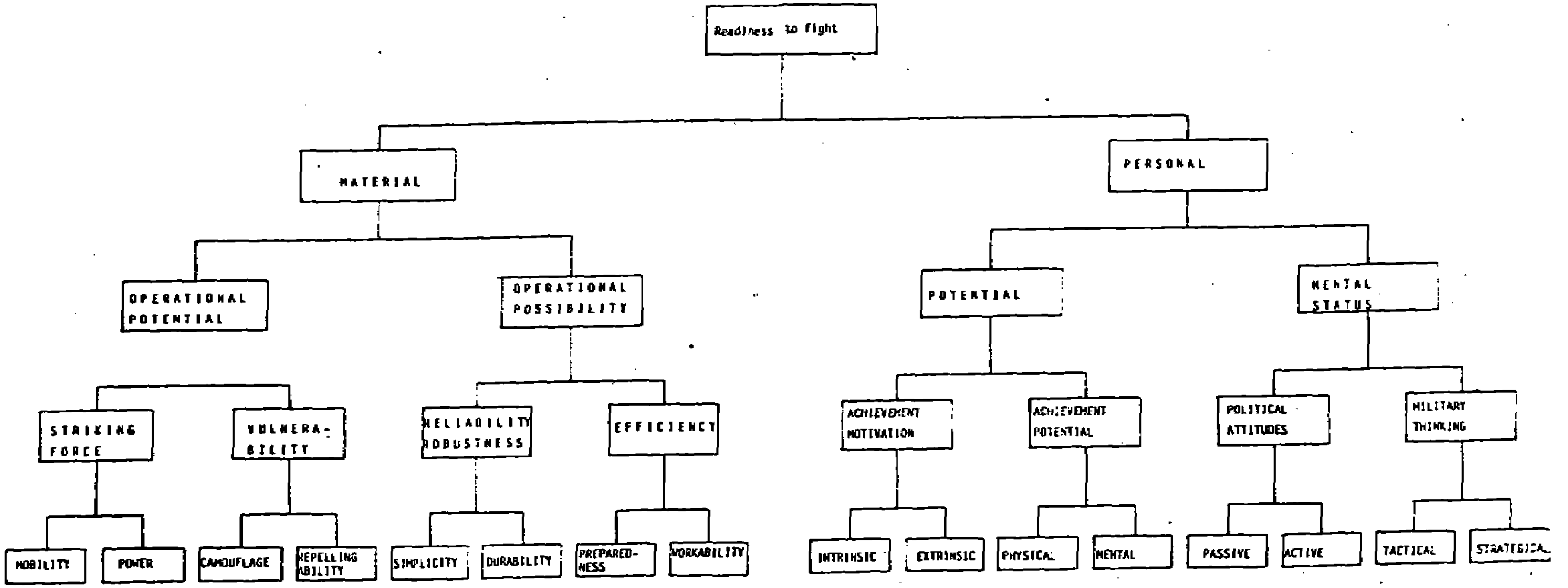


Figure 3: Measure of Effectiveness (here readiness to fight)

III. Vague relationship

Statements relating phenomina to each other can also be vague. Examples of this type of statements are "not much larger than", "approximately equal to..." etc.

Intrinsic and informational vaguenesses are best characterized by the following citations:

In 1923 B. Russell noted already:
"All traditional logic habitually assumes that precise symbols are being employed. It is therefore not applicable to this terrestrial life but only to an imagined celestial existence."

Principle of Incompatibility: (Zadeh 1965): "As the complexity of a system increases our ability to make precise and yet significant statements about its behaviour diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics."

2. What are Fuzzy Sets

The notion of a fuzzy set and the axiomatic system of fuzzy set theory was put forward by L. Zadeh in 1965. Here just some of his basic definitions:

Fuzzy Set:

If $X = \{x\}$ is a collection of objects denoted generically by x then a Fuzzy Set A in X is a set of ordered pairs

$$A = \{(x, \mu_A(x)) \mid x \in X\} .$$

$\mu_A(x)$ is called a membership function or grade of membership of x in A which maps X to the membership space M. The range of the membership function is a subset of the nonnegative real numbers whose supremum is finite.

Equality: Two fuzzy sets A and B are equal iff

$$\mu_A(x) = \mu_B(x) \qquad \text{for all } x \varepsilon X$$

Intersection: The membership function of $A \cap B$ is given by

("and")

$$\mu_{A \cap B}(x) = \text{Min } (\mu_A(x), \mu_B(x))$$

Union: The membership function of $A \cup B$ is defined as

("or")

$$\mu_{A \cup B}(x) = \text{Max } (\mu_A(x), \mu_B(x))$$

Example: Let X = {10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110} be possible speeds (mph) at which cars can cruise over long distances. The fuzzy set A of "comfortable speeds for long distances" may be defined by certain individual as

A = {(30, 0.7), (40, 0.75), (50, 0.8), (60, 0.8), (70, 1.0)
(80, 0.8), (90, 0.3)}

Support: The support of a fuzzy set A is a set S(A) such that $x \varepsilon X(A)$ iff $\mu_1(x) > 0$.

Normality: It has already been mentioned that the membership function is not limited to values between 0 and 1. If $\text{Sup}_X \mu_A(x) = 1$ the fuzzy set A is called normal. A non-empty fuzzy set A can always be normalized by dividing $\mu_A(x)$ by $\text{Sup}_X \mu_1(x)$.

Algebraic Product: The membership function μ_{AB} of the algebraic product of two fuzzy sets A and B is defined as:

$$\mu_{AB} = (\mu_A \cdot \mu_B),$$

Algebraic Sum: The membership function of the algebraic sum of A and B is also defined by its membership function:

$$\mu_{A \oplus B} = \mu_A + \mu_B - \mu_A \mu_B$$

Relation: A fuzzy relation, R, in the product space $X \times Y = \{(x,y)\} \mid x \in X$, $y \in Y$ is a fuzzy set in $X \times Y$, whose membership function μ_R assoziates with each ordered pair (x,y) a grade of membership $\mu_R(x,y)$ in R. An n-ary Relation in a product space $X = X^1 \times X^2 \times \ldots \times X^n$ is then characterized by a corresponding n-variate membership function.

The notion of a "decision" has always had very many different semantic interpretations. Two distinct approaches are of particular importance: In cognitive (descriptive) decision theory a decision is an information processing process which can either lead to an evaluation (measure of effectivness) to a ranking of different alternatives or to an "optimal solution". Probably better known is the definition of a decision which is used in normative decision theory (logic of decisions). Here a decision is the act of selecting a specific solution (action) which is feasable (element of the solution space) and optimal (f.i. maximising an objective function).

The latter notion lead Bellman and Zadeh in 1970 to define a decision in a fuzzy environment as follows:

In a fuzzy decision situation the constraints as well as the objective function(s) can be fuzzy sets, characterized by their membershipfunctions and the "decision" is the (fuzzy) set of all activities which are members of the fuzzy constraints sets and the fuzzy sets characterizing the objective function(s) i.e. the "decision" is the intersection of all fuzzy sets involved (objective functions and constraints).

Example 1:

The board of directorsis trying to find the "optimal" dividend to be payed to the shareholders. For financial reasons it ought to be attractive and for reasons of wage negotiations it should be modest. The fuzzy set of the objective function "attractive dividends" could for instance be defined by:

$$\mu_0(x) = \begin{cases} 1 & \text{for } x \geq 5.8 \\ \frac{1}{100.000}\,(-464x^3 + 7961x^2 - 14530x + 7033) & \text{for } 1 \leq x \leq 5.8 \\ 0 & \text{for } x \leq 1 \end{cases}$$

and the fuzzy set (constraint) "modest dividends" by

$$\mu_C(x) = \begin{cases} 1 & \text{for } x \leq 1.2 \\ \frac{1}{100.000}\,(3270x^3 - 31805x^2 + 62206x + 65501) & \text{for } 1.2 \leq x \leq 6 \\ 0 & \text{for } x \geq 6. \end{cases}$$

The fuzzy set "decision" is then $\mu_D = \text{Min}\,(\mu_0(x), \mu_C(x))$

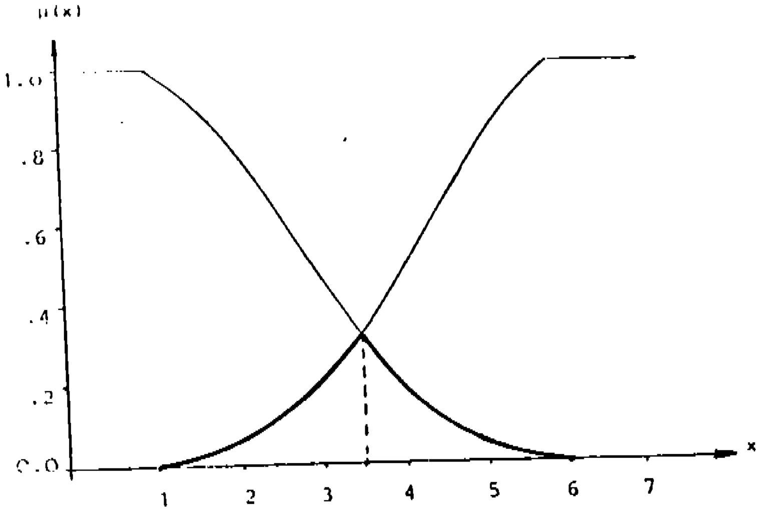


Figure 4: A fuzzy decision

So far the "decision" is still a fuzzy set. A reasonable way of picking a dividend would now be to select the solution which has the highest degree of membership in the decision set. In our example this "optimal decision" would be the dividend x^0 = 3.5% with a degree of membership in the decision set of $\mu_D(x^0) = .338$.

A second example which could also be regarded as a decision is the following:

Example 2:

An instructor at a university has to decide how to grade written test papers. Let us assume that the problem to be solved in the test was a linear programming problem and that the student was free to solve it either graphically or using the simplex method. The student has done both. The student's performance is expressed - for the graphical solution as well as for the algebraic solution - as the achieved degree of membership in the fuzzy sets 'good graphical solution' (G) and 'good simplex solution' (S), respectively. Let us assume that he reaches

$$\mu_G = 0.9 \qquad \text{and} \qquad \mu_S = 0.7.$$

If the grade to be awarded by the instructor corresponds to the degree of membership of the fuzzy set 'good solutions of linear programming problems' it would be quite conceivable that this grade μ_{LP} could be determined by

$$\mu_{LP} = \text{Max}(\ \mu_G,\ \mu_S) = \text{Max}\ (0.9,\ 0.7) = 0.9.$$

3. Fuzzy Optimization

As a special structure of an optimization problem we shall choose Linear Programming. Let us consider the following cases:

A company wanted to decide on the size and structure of its truck fleet. Four differently sized trucks (x_1 through x_4) were considered. The objective was to minimize cost and the constraints were to supply all customers (who have a strong seasonally fluctuating demand). That meant: certain quantities had to be moved (quantity constraint) and a minmum number of customers per day had to be contacted (routing constraint). For other reasons, it was required that at least 6 of the smallest trucks be included in the fleet. The management wanted to use quantitative analysis and agreed to the following suggested linear programming-approach:

Minimize

$$41.400x_1 + 44.300x_2 + 48.100x_4 + 49.100x_4$$

subject to constraints

$$0.84x_1 + 1.44x_2 + 2.16x_3 + 2.40x_4 \geq 170$$

$$16x_1 + 16x_2 + 16x_3 + 16x_4 \geq 1.300$$

$$x_j \geq 6.$$

The solution was $x_1 = 6$, $x_2 = 17.85$, $x_3 = 0$, $x_4 = 58,64$. Min Cost = 3.670.850. Since the management felt that is was forced into giving precise constraints (because of the model) in spite of the fact that it would rather have given some intervals, the following "fuzzy" approach was used:

Starting from the problem:

$$\begin{array}{ll} \text{Minimize} & Z = cx \\ \text{subject to constraints} & Ax \leq b \\ & x \geq 0, \end{array}$$

the adopted "fuzzy" version was

$$\begin{aligned} cx &\underset{\sim}{\leq} Z \\ Ax &\underset{\sim}{\leq} b \\ x &\geq 0. \end{aligned}$$

Here c is the vector of coefficients of the objective function, b is the vector of constraints (in our case the amount to be shipped, the number of customers to be contacted and the minimum number of small trucks required), and A is the coefficient matrix. The symbol "$\underset{\sim}{<}$" denotes the fuzzied version of "$\leq$" and reads "essentially smaller than or equal to".

We now define a function f: $R^{m+1} \rightarrow [0,1]$ such that

$$f(Ax) = \begin{cases} 0 \text{ if } Ax \leq b \text{ and } cx \leq Z \text{ is strongly violated} \\ \\ 1 \text{ if } Ax \leq b \text{ and } cx \leq Z \text{ is satisfied.} \end{cases}$$

Using the simplest version of the function (Ax,cx) we assume it to be linear and the intersection of the (fuzzy) constraints and the (fuzzy) objective function.

Thus

$$f(Ax,cx) = f(Bx) = \text{Min } f_i(Bx)_i), \; x \geq 0$$

with

$$f_i(Bx)_i) = \begin{cases} 1 & \text{for } (Bx)_i \leq b_i \\ 1 - \dfrac{Bx_i - b_i}{d_i} & \text{for } b_i < Bx_i \leq b_i \quad + d_i \\ 0 & \text{for } (Bx)_i > b_i + d_i \end{cases}$$

where d_i are subjectively chosen constants of admissible violations of the constraints,

$f_i(Bx)_i$ is the membership function of the i-th row of the linear system Bx_i,

Min $f_i(Bx)_i)$
is the "fuzzy" decision

and

Max Min $f_i(Bx)_i)$

the decision with the highest degree of membership.

Substituting

$$b_i' = \frac{b_i}{d_i}$$

$$B_i' = \frac{B_i}{d_i}$$

componentwise and simplifying it by dropping the "1" (which does not change the problem!) we arrive at the following problem:

$$\max_{x \geq 0} \min_i \left(b_i' - (B'x)_i\right) \qquad (x)$$

or

$$\max_{x \geq 0} \mu_D(x)$$

As is well known, this problem is equivalent to solving the following linear program:

Maximize λ (xx)

subject to constraints $\lambda \leq b_i - (Bx)_i$, $i = 0,1,..,m$

$x \geq 0$

The optimal solution to (x) is also the optimal solution to (xx).

When this approach was applied to our problem the following assumptions were made:

1) Total cost should not rise above 4.200.000 (budget limit).
2) The "unfuzzy" constraints are minimum requirements and management would feel much better if there was some "leeway".
3) The linear approximation of the membership functions are acceptable.
4) There are no interdepencies between the constraints.
5) Weighting of the constraints is taken care of by defining the constants d_i.

The results are shown in the following table:

Unfuzzy	Fuzzy	
	$\mu = 0$	$\mu = 1$
Objective Function	4.200.000	3.700.000
1st constraint 170	170	180
2nd constraint 1.300	1.300	1.400
3rd constraint 6	6	12

Our non-fuzzy equivalent problem in the form of (xx) is then:

Maximize λ

subject to constraints

$$\lambda \leq 7.4 - 0.083x_1 - 0.089x_2 - 0.096x_3 - 0.098x_4$$

$$\lambda \leq - 18 + 0.084x_1 + 0.144x_2 + 0.216x_3 + 0.24x_4$$

$$\lambda \leq - 14 + 0.16x_1 + 0.16x_2 + 0.16x_3 + 0.16x_4$$

$$\lambda \leq - 2 + 0.167$$

$x_1, x_2, x_3, x_4 \geq 0.$

Solution:

Unfuzzy	Fuzzy
$x_1 = 6$	$x_1 = 17.41$
$x_2 = 17.85$	$x_4 = 66.54$
$x_4 = 58.65$	
$Z = 3.918.850$	$Z = 3.988.257$
Constraints:	
1. 171.5	174.2
2. 1.320	1.342.4
3. 6	17.4

As can be seen from the solution, "leeway" has been provided with respect to all constraints and at additional cost of 1.7 percent.

The main advantage, compared to the unfuzzy problem formulation, is the fact that the decision maker is not forced into a precise formulation because of mathematical reasons even though he might only be able or willing to describe his problem in fuzzy terms. Linear membership functions are obviously only a very rough approximation. Membership functions which monotonically increase or decrease, respectively, in the interval of $[b_i, b_i+d_i]$ can also be handled quite easily.

The above fuzzy LP approach can also be used to tackle multi-criteria-problems very efficiently.

In the following we shall restrict our considerations to linear programming problems with vectorvalued objective functions. For instance we shall use the following example:

Example 3

$$\text{"max" } Z(x) = \begin{pmatrix} -1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{aligned} \text{s.th. } -x_1 + 3x_2 &\leq 21 \\ x_1 + 3x_2 &\leq 27 \\ 4x_1 + 3x_2 &\leq 45 \\ 3x_1 + x_2 &\leq 30 \\ x_1, x_2 &\geq 0. \end{aligned}$$

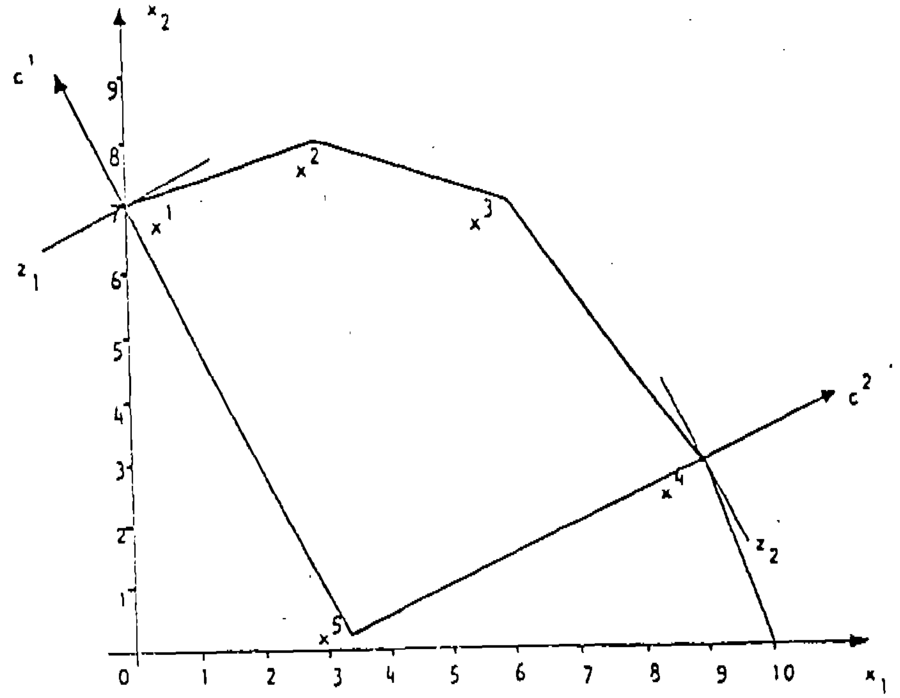


Figure 5: A Vector Maximum Problem

Figure 5 shows the solution space of this problem. The "complete solution" is the edge $x^1 - x^2 - x^3 - x^4$. x^1 is optimal with respect to objective Function $z_1 = -x_1 + 2x_2$. x^4 is optimal with respect to objective function $z_2 = 2x_1 + x_2$. The "optimal" values are $z_1\ (x^1) = 14$ and $z_2\ (x^4) = 21$, respectively. For $x^1 = (7,0)$ $z_2 = 7$ and $x^4 = (9,3)$ yields $z_1 = -3$. Solution $x^5 = (3.4,\ 0.2)$ is the solution which yields $z_1 = -3$, $z_2 = 7$ the lowest "justifyable" values of the objective functions in the sense that a ? further decrease of the value of one objective functions could not be "balanced" or even counteracted by an increase of the value of the other objective function.

We shall now apply our FLP approach to the vectormaximum problem and make the following assumptions:
The membership functions $\mu_1\ (x)$ and $\mu_2\ (x)$ of the fuzzy sets characterizing the objective functions rise linearly from 0 to 1 at the highest achievable values of $z_1 = 14$ and $z_2 = 21$, respectively. Thus

$$\mu_1\ (x) = \begin{cases} 0 & \text{for } z_1 \leq -3 \\ \dfrac{z_1(x) + 3}{17} & \text{for } -3 < z_1 \leq 14 \\ 1 & \text{for } 14 < z_1 \end{cases}$$

$$\mu_2(x) = \begin{cases} 0 & \text{for } z_2 \leq 7 \\ \dfrac{z_2(x) - 7}{14} & \text{for } 7 < z_2 \leq 21 \\ 1 & \text{for } 21 < z_2 \end{cases}$$

We arrive at the following problem

Max λ

$$\begin{array}{lrl} \text{s.th.} & \lambda \leq & -0.05882x_1 + 0.117\ x_2 + 0.1764 \\ & \lambda \leq & +0.1429\ x_1 + 0.0714x_2 + 0.5 \\ & 21 \geq & -x_1 + 3x_2 \\ & 27 \geq & x_1 + 3x_2 \\ & 45 \geq & 4x_1 + 3x_2 \\ & 30 \geq & 3x_1 + x_2 \\ & x \geq 0 & \end{array}$$

depicted in Figure 6.

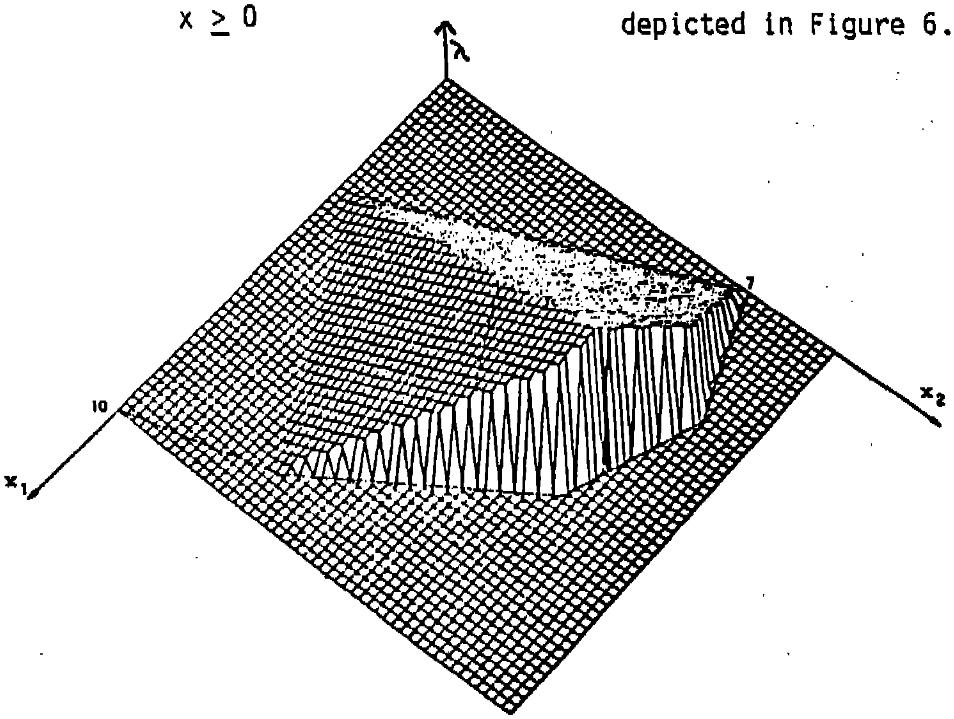


Figure 6: The Vector Maximum Problem

Nonlinear membership functions

Hersh and Caramazzy have shown that membership functions of the following type exist: (also compatible with economic laws of decreasing rate of increase of utility):

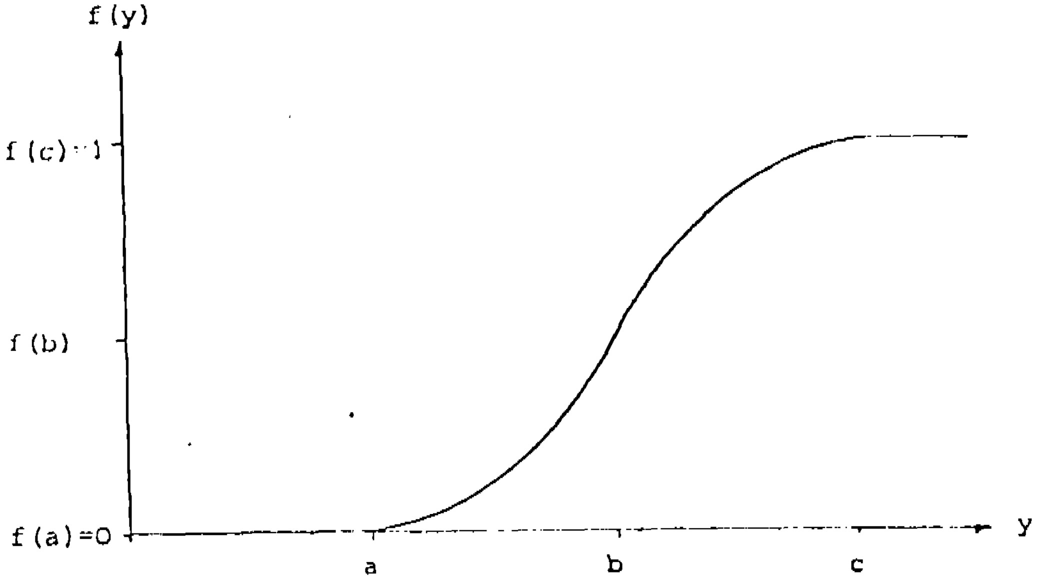


Figure 7: Nonlinear Membership Function

Such a nonlinear membership function can be described by a hyperbolic function. For each objective function z_j, $j = 1,\dots,k$, the corresponding hyperbolic membership function μ^H is defined as follows.

$$\mu^H_{z_j}(x) := \frac{1}{2}\,\frac{e^{(z_j(x)-\frac{z_j^m+z_j^o}{2})\alpha_j} - e^{-(z_j(x)-\frac{z_j^m+z_j^o}{2})\alpha_j}}{e^{(z_j(x)-\frac{z_j^m+z_j^o}{2})\alpha_j} + e^{-(z_j(x)-\frac{z_j^m+z_j^o}{2})\alpha_j}} + \frac{1}{2},$$

where α_j is a parameter.

The following figures illustrate the hyperbolic membership functions of the objective functions z_1 and z_2 of the given example.

With $\alpha_j = 3 \,/\, \frac{1}{2}\,(z_j^0 - z_j^m)$ we get

$$\mu_{z_1}^h(x) = \frac{1}{2}\,\frac{e^{(z_1(x) - 5.5)\frac{6}{17}} - e^{-(z_1(x) - 5.5)\frac{6}{17}}}{e^{(z_1(x) - 5.5)\frac{6}{17}} + e^{-(z_1(x) - 5.5)_{17}}} + \frac{1}{2}$$

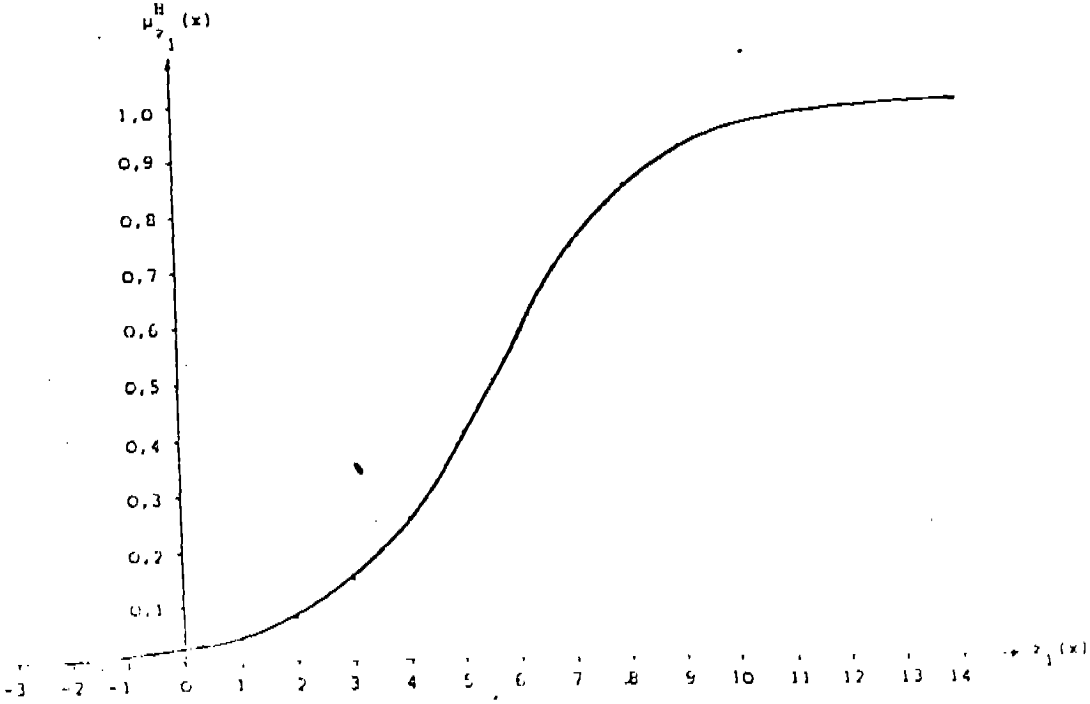


Figure 9: Nonlinear Membership Function for Example

$$\mu_{z_2}^h(x) = \frac{1}{2}\,\frac{e^{(z_2(x) - 14)\frac{6}{14}} - e^{-(z_2(x) - 14)\frac{6}{14}}}{e^{(z_2(x) - 14)\frac{6}{14}} + e^{-(z_2(x) - 14)\frac{6}{14}}} + \frac{1}{2}$$

The problem now reads:

Max λ

such that

$$\lambda - \frac{1}{2}\,\frac{e^{(z_j(x) - \frac{z_j^m + z_j^o}{2})\alpha_j} - e^{-(z_j(x) - \frac{z_j^m + z_j^o}{2})\alpha_j}}{e^{(z_j(x) - \frac{z_j^m + z_j^o}{2})\alpha_j} + e^{-(z_j(x) - \frac{z_j^m + z_j^o}{2})\alpha_j}} \leq \frac{1}{2}, \quad j = 1, \ldots, k$$

$$Ax \leq b$$

$$x \geq 0$$

$$\lambda \geq 0.$$

This is a nonlinear programming problem with one linear objective function, k nonlinear and m+n+1 linear restrictions. We shall now show that there exists an equivalent linear optimization problem.

It can be shown (Leberling 1980) that for

$x_{n+1} = \tanh^{-1}(2\lambda - 1)$

an equivalent formulation is

Max $\quad x_{n+1}$

such that $\quad \alpha_j z_j(x) - x_{n+1} \geq \frac{1}{2}\alpha_j(z_j^m + z_j^0), \; j = 1, \ldots, k.$

$$Ax \leq b$$

$$x \geq 0.$$

For our problem: (For $\alpha_j = 3/\,\frac{1}{2}\,(z_j^0 - z_j^m)$) we obtain the following problem formulation:

Max λ

such that

$$\lambda - \frac{1}{2}\,\frac{e^{(-x_1+2x_2-5,5)\frac{6}{17}} - e^{-(-x_1+2x_2-5,5)\frac{6}{17}}}{e^{(-x_1+2x_2-5,5)\frac{6}{17}} + e^{-(-x_1+2x_2-5,5)\frac{6}{17}}} \leq \frac{1}{2}$$

$$\lambda - \frac{1}{2}\,\frac{e^{(2x_1+x_2-14)\frac{6}{14}} - e^{-(2x_1+x_2-14)\frac{6}{14}}}{e^{(2x_1+x_2-14)\frac{6}{14}} + e^{-(2x_1+x_2-14)\frac{6}{14}}} \leq \frac{1}{2}$$

$$\begin{array}{rcrl}
-x_1 & + & 3x_2 & \leq 21 \\
x_1 & + & 3x_2 & \leq 27 \\
4x_1 & + & 3x_2 & \leq 45 \\
3x_1 & + & x_2. & \leq 30 \\
x_1 & & & \geq 0 \\
 & & x_2 & \geq 0 \\
 & & \lambda & \geq 0.
\end{array}$$

The equivalent problem formulation is

Max x_3

such that

$$\begin{array}{rcrcrl}
-6x_1 & + & 12x_2 & - & 17x_3 & \geq 33 \\
12x_1 & + & 6x_2 & - & 14x_3 & \geq 84 \\
-x_1 & + & 3x_2 & & & \leq 21 \\
x_1 & + & 3x_2 & & & \leq 27 \\
4x_1 & + & 3x_2 & & & \leq 45 \\
3x_1 & + & x_2 & & & \leq 30 \\
x_1 & & & & & \geq 0 \\
 & & x_2 & & & \geq 0.
\end{array}$$

The optimal solution is

$(x_1^{opt}; x_2^{opt}; x_3^{opt})^T = (5.03; 7.32; 1.45)^T$.

The maximum degree of membership of satisfaction

$\lambda\, opt = \frac{1}{2} \tanh (x_3^{opt}) + \frac{1}{2} = 0.95$

is achieved at the solution $(x_1^{opt}; x_2^{opt})^T = (5.03; 7.32)^T$.
This solution os the wanted compromise solution of the given LVOP.

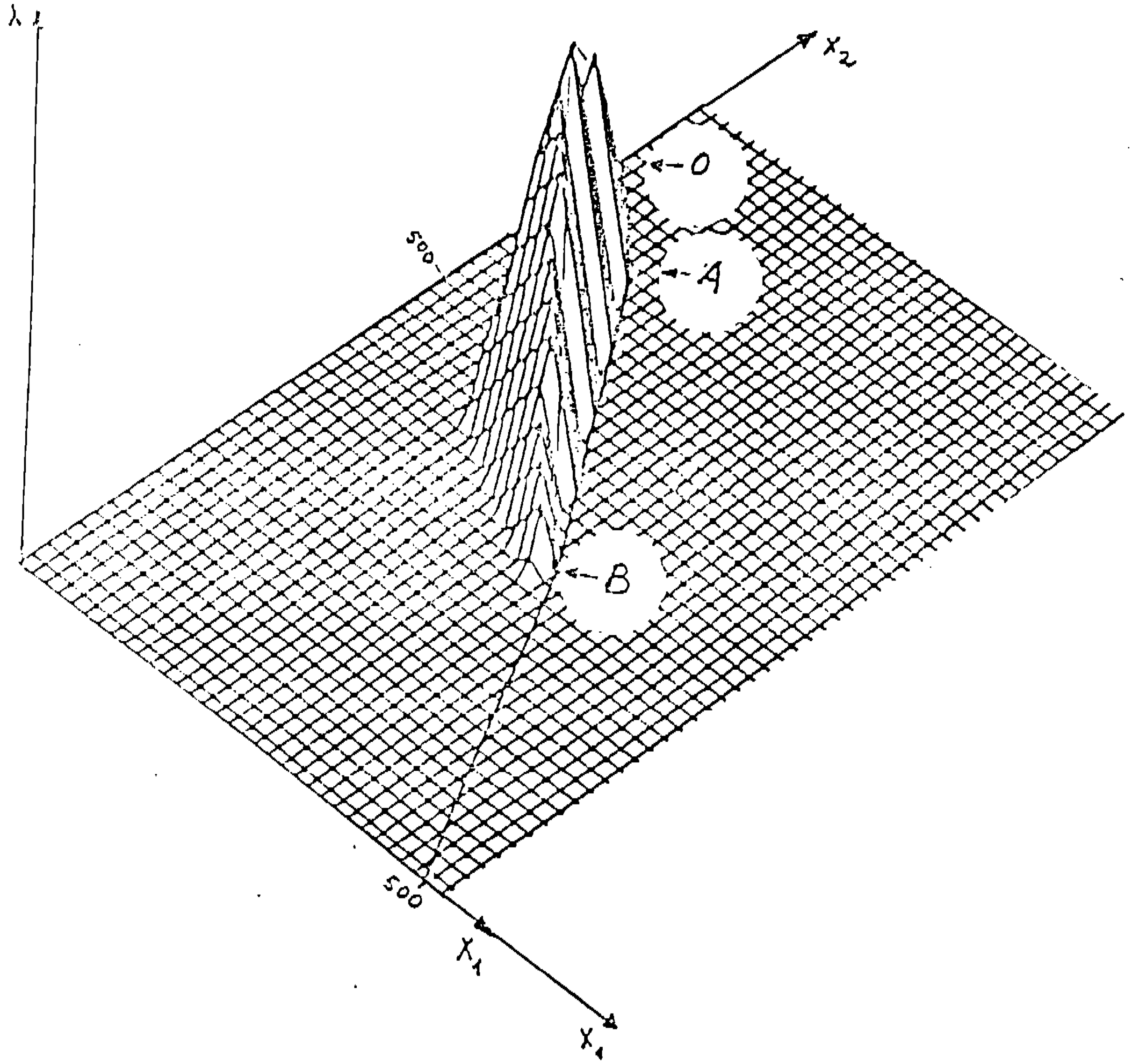


Figure 9: The Nonlinear Case

4. Fuzzy Evaluation

Examples of evaluation problems are the determination of creditworthiness and "readiness to fight", which were mentioned at the beginning. Evaluation and optimization problems certainly have one feature in common: Fuzzy sets representing subjective categories, objectives or constraints have to be aggregated.

Comparing the dividend example with the grading problem it seems appropriate to reconsider the original definition of a decision as the "intersection of all fuzzy sets involved". The grading example would suggest rather to use the union as a model and the two evaluation models do not give us any hint in which way the aggregation ought to be accomplished.

The two definitions of decisions - as the intersection or the union of fuzzy sets - imply essentially the following:

The interpretations of a decision as the intersection of fuzzy sets implies no positive compensation (trade-off) between the degrees of membership of the fuzzy sets in question, if either the minimum or the product is used as an operator. Each of them yields degrees of membership of the resulting fuzzy set (decision) which are on or below the lowest degree of membership of all intersecting fuzzy sets (see Example 1).

The interpretation of a decision as the union of fuzzy sets, using the max-operator, leads to the maximum degree of membership achieved by any of the fuzzy sets representing objectives or constraints. This amounts to a full compensation of lower degrees of membership by the maximum degree of membership (see Example 2).

Observing managerial or military decisions one finds that there are hardly any decisions with no compensation (trade-offs) between either different degrees of goal achievement or the degree to which restrictions are limiting the scope of decisions. The compensation, however, rarely ever seems to be "complete" such as would be assumed using the max-operator.

It may be argued that compensatory tendencies in human aggregation are responsible for the failure of some classical operators (min, product, max) in empirical investigations.

It is crucial for the appropriate modelling of decisions of whatever form to know:

1. How to model properly subjective categories by fuzzy sets (i.e. how to determine membership functions).
2. How to aggregate fuzzy sets appropriately in order to arrive at a decision, a judgment, a ranking or an evaluation.

Two ways of proceeding are conceivable:

a) The axiomatic approach (as in utility theory), establishing an axiomatic system which is at least plausible and derive the resulting mathematical models for aggregators and membership functions.

b) Empirical investigations into the ways people think of subjective categories (i.e. determining appropriate membership functions) and aggregate them to arrive at conclusions (i.e. determining models for aggregatos).

Some authors have used the first approach [2] . We shall report on some results of doing research of the latter type.

5. Empirical Results

5.1 Aggregators (Connectives)

In earlier studies [6,9] it was shown, that neither minimum nor product operator model properly the human "logical and". During these investigations it appeared, however, that in managerial decisions the "logical and" (without trade-offs") is not used at all.

Our hypothesis was that human beings use many non-verbal connectives in their thinking and reasoning. One type of these connectives may be called "merging connectives" which may be represented by the "compensatory and". Being forced to verbalize them men possibly map the set of "merging connectives" into the set of the corresponding language connectives ("and", "or"). Hence, when talking, they use the verbal connectives which they feel closest to their "real" non-verbal connective.

Thus a new word "compensatory and" had to be coined and possible mathematical models for it tested. A number of models, such as the minimum, the maximum, the geometric mean, the arithmetic mean etc., were tested. In addition a special model, the so-called γ-operator was tested. The following figures indicate some of the results :

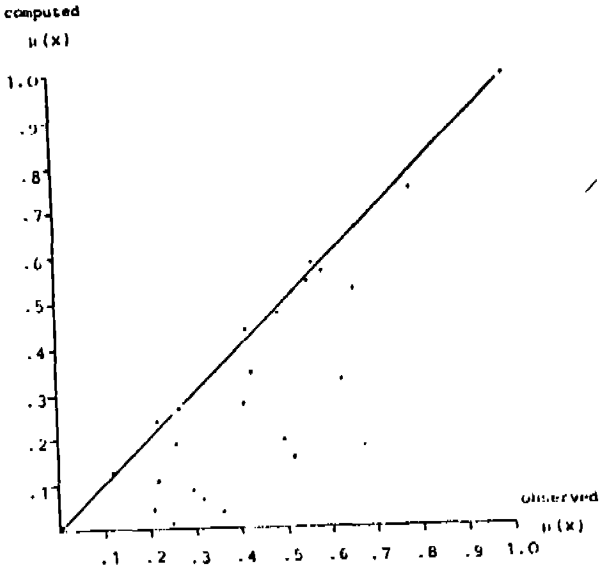
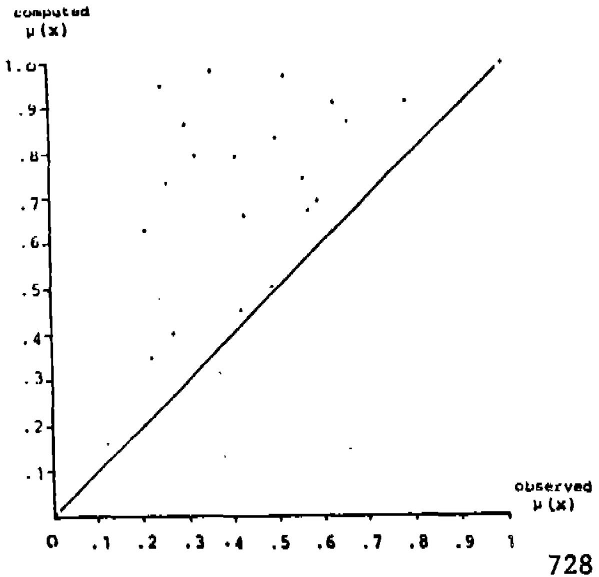


Figure 10: Min-operator: Observed versus computed grades of membership



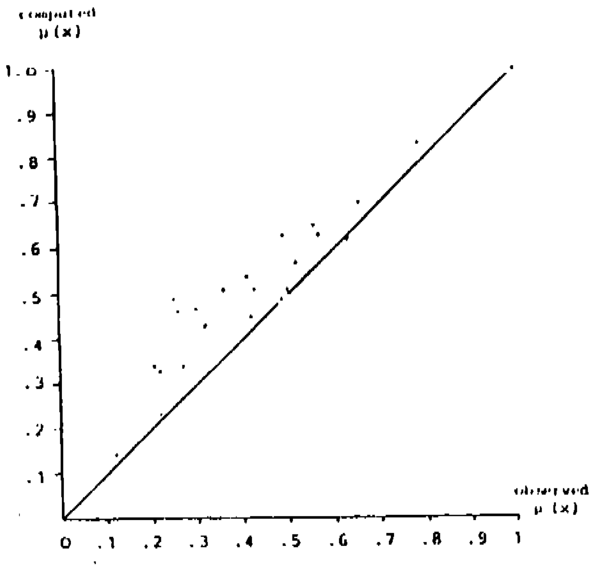


Figure 12: Arithmetic mean: Observed versus computed grades of membership

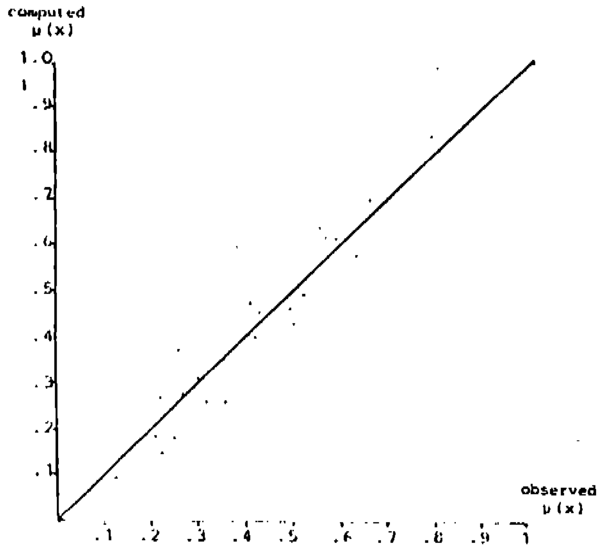


Figure 13: γ-operator: Observed versus computed grades of membership

The γ-operator performed best. The idea behind it is the following:

If several operators are necessary in order to describe a variety of phenomena, the question arises, how many operators are needed, as each important situation in practise would then call for an adequate model. Moreover, one would be forced to assume that man has a decision rule enabling him to choose the right connective for each situation. The pursuit of this train of thought and especially its application implies a lot of difficulties. We feel that one way to bypass these difficulties is to generalize the classical concept of connectives by introducing a parameter γ which may be interpreted as "grade of compensation".

Each point of the continuum between "and" and "or" represents a different operator. One way to formalize this idea is to find an algebraic representation for a weighted combination of the non-compensatory "and" and the fully compensatory "or". The more there is a tendency for compensation the more the "or" becomes effective and vice versa. As "extremal" operators we prefer the product and the algebraic sum. Of course, other "extremes" are conceivable, for instance minimum and maximum. But in our opinion these models are handicapped as they do not reflect the interaction of membership values. Thus we define:

$$\mu_{A\theta B} = \mu_{A\cap B}^{1-\gamma} \cdot \mu_{A\cup B}^{\gamma} \qquad (x)$$

The membership of an object in the set $A\theta B$ equals the product of the weighted membership values for the intersection and the union.
If the intersection and the union are algebraically represented by the product and the algebraic sum, respectively, then the operator becomes:

$$\mu_{\theta} = \left(\prod_{i=1}^{m} \mu_i\right)^{1-\gamma} \left(1 - \prod_{i=1}^{m} (1 - \mu_i)\right)^{\gamma}, \quad 0 \leq \mu \leq 1, 0 \leq \gamma < 1$$

$i = 1,2,\ldots,m$, m = number of sets to be connected.

It can be shown that the γ-operator is pointwise injective, continuous, monotonous, commutative and in accordance with the truth tables of dual logic.

Using this operator a model with very high predictive power could, for instance, be derived for the credit worthiness decision. The γ's were determined from observations and then the following hierarchy resulted:

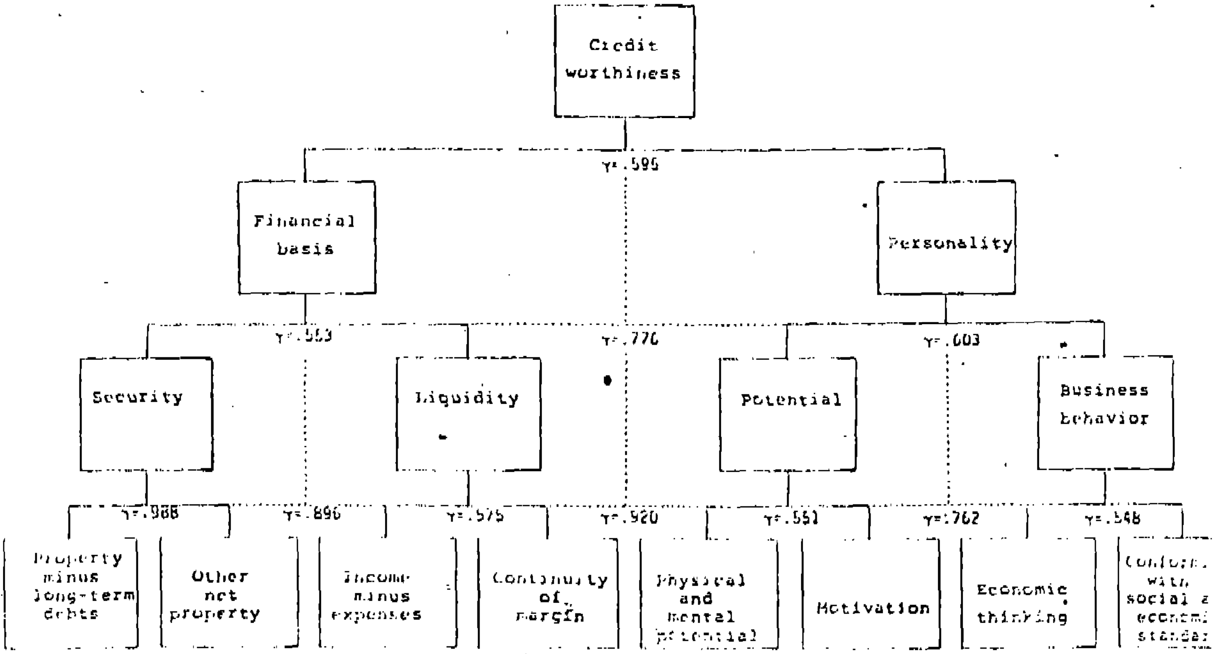


Figure 14: Empirically determined values of γ for the creditworthiness hierarchy

5.2 Membership Functions

When determining membership functions empirically one first has to decide on which scale level the resulting function has to be. This depends on the intended use of membership functions.If they are to be used, for instance, in the framwork of fuzzy linear programming then they have to be on an absolute scal level. It is well known, however, that the human being is no reliable measurement device on this level. In [6] it was shown how to generate "quasi-cardinal" degrees of membership from human answers. These degrees of membership did not constitute, however, "membership functions"(membership as a function of a variable which is to be optimized) as one needs them for mathematical programming.

For this purpose another approach was used:

Given that for a subjective category under consideration (for instance "young men") there exists a judgmental scale (years), and an evaluational scale (degree of being young) then the membership function of the fuzzy set can be viewed as the "distance" of these two scales. If the distance function is called d(x) then the membership function (Type II) is

$$\mu(x) = \frac{1}{1 + d(x)}$$

Different models for the distance function are possible. The one that performed best so far is - not surprisingly - the logistic function $d(x) = e^{-a(x+b)}$.

The following pictures indicate some of the results of empirically testing the membership function

$$\mu(x) = \frac{1}{1 + e^{-a(x+b)}} \ .$$

The lines represent the mathematical model after the parameters a and b have been determined for proper calibration. The dots represent the observations.

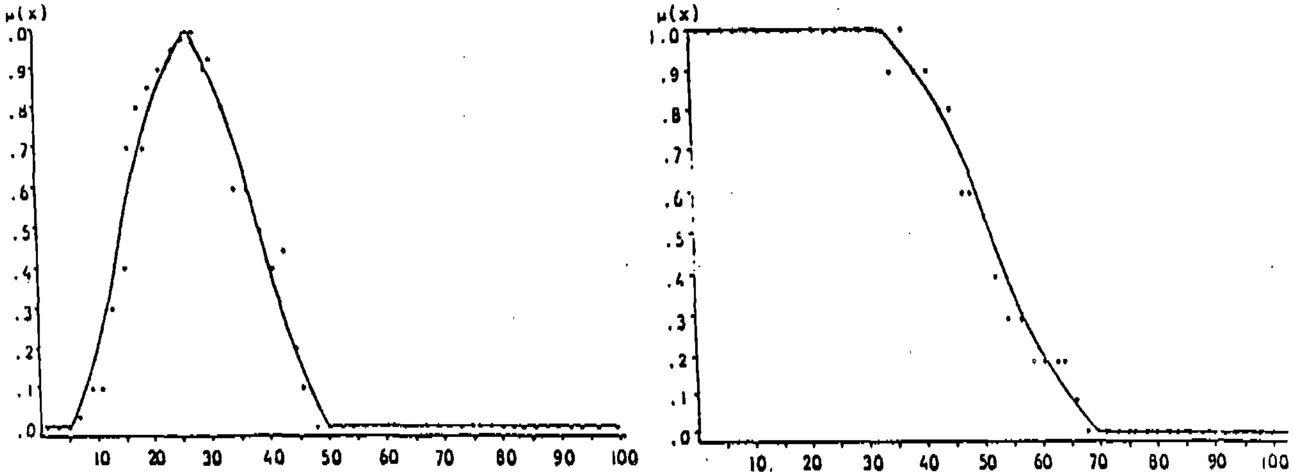


Figure 15: Fuzzy Set "Young Men"

References:

1. Bellman, R.E. and Zadeh, L.: Decision Making in a Fuzzy Environment. Mgt. Sc. 17 (1970) 141-164.

2. Hamacher, H.: Über logische Verknüpfungen unscharfer Aussagen und deren zugehörige Bewertungsfunktionen. In: R. Trappl, G.J. Klir, L. Ricciardi (eds.). Progress in Cybernetics and Systems Research, New York 1978.

3. Hersh, H.M. and Caramazza, A.A.: A Fuzzy Set Approach to Modifiers and Vagueness in Natural Language. J. Exp. Psych. 105 (1976), 254-276.

4. Kwakernaak, H. and Baas, S.M.: Rating and Ranking of Multi-Aspect Alternatives using Fuzzy Sets. In: Automatica 13 (1977), 47-58.

5. Leberling, H.: On Finding Compromise Solutions in Multicriteria Problems using the Fuzzy Min-operator.

6. Thole, U., Zimmermann, H.-J. and Zysno, P.: On the suitability of Minimum and Product Operators for the Intersection of Fuzzy Sets. FSS 2 (1979) 167-180.

7. Zadeh, L.A.: Fuzzy Sets. Information and Control 8 (1965) 338-353.

8. Zimmermann, H.-J.: Fuzzy Programming and Linear Programming with Several Objective Functions. FSS 1 (1978), 45-55.

9. Zimmermann, H.-J. and Zysno, P.: Latent Connectives in Human Decision Making. FSS 4 (1980), 37-51.

ARTILLERY CONTROL ENVIRONMENT: AN EXPERIMENTAL TOOL

Jill H. Smith and Jock O. Grynovicki
Experimental Design and Analysis Branch/ACE Team
Systems Engineering and Concepts Analysis Division
Ballistic Research Laboratory
Armament Research and Development Center/USAAMCCOM
Aberdeen Proving Ground, MD 21005

ABSTRACT. The Army is fielding a new digital communications system, the TACFIRE system, shown for the brigade-area in Figure 1. In order to investigate the command, control, and communications issues associated with the new devices, the Artillery Control Environment (ACE) was developed. ACE is a real-time, multiplayer, interactive simulation system run on a commercial computer that interfaces with the tactical equipment through a bit box (microprocessor). This paper will discuss the preparations, experimental design, data collection and analysis methods for the first experiment with military players interfaced with the Artillery Control Environment software to be conducted 8 May - 10 June 83.

I. INTRODUCTION. In May 1982, the HELBAT (Human Engineering Laboratory Battalion Artillery Test) Executive Committee agreed that the Ballistic Research Laboratory Artillery Control Environment (ACE) and HELBAT activities should be combined to develop a Command Post Exercise Research Facility (CPXRF). The CPXRF can be used to demonstrate the use of commercial automatic data processing (ADP) technology for the RDT&E (research, development, testing, and evaluation) of tactical ADP fire support control systems and also, for training personnel to operate these systems; the HEL/BRL CPXRF, however, will primarily be used for research and exploratory development work. Further, an ACE/CPXRF Subcommittee was formed to provide joint DARCOM-TRADOC guidance in the development of ACE technology and use of the CPXRF. The ACE software is a key tool in the CPXRF. The software features the ability to automatically load live players with messages produced by target acquisition and fire direction simulators while recording all the message traffic that flows between the live and simulated players.

An overview of the CPX Research Facility and ACE program is given in the 1982 Sept-Oct issue of the Field Artillery Journal in an article "HELBAT/ACE Fire Support Control Research Facility" by Mr. Barry Reichard. The layout of the facility is shown in Figure 2.

II. PURPOSE. The experiment detailed in this plan is the first test in which military players will be interfaced with the Artillery Control Environment (ACE) software. The purpose of this experiment is to demonstrate the feasibility of using the automated techniques of the CPX Research Facility for fire support control experiments.

To demonstrate this capability, it was decided that a study of the effects of message intensity and degraded communication on a Fire Support Team Headquarter's (FIST HQ) ability to perform fire support coordination would be appropriate. Message intensity is defined to be a function of message type, message rate, and message content.

III. TEST CONCEPT.

A. Objectives

1) To determine the effect of message intensity on the FIST HQ's ability to perform fire support coordination.

2) To determine the effect of degraded communication on the FIST HQ's ability to perform fire support coordination.

3) To determine if message intensity and degraded communication have a combined effect on fire support coordination.

B. Measures of Performance

A measure of performance (MOP) is a response that is used to quantify the effects of the factors to be evaluated. Because all of our objectives investigate the effect on fire support coordination, the measures of performance will be the same for all three objectives. The following measures of performance will be computed on each two hour cell of the test:

1) Number of messages serviced (i.e. for which a response has been generated)/total number of messages received by the FIST DMD (digital message device)

2) Time required to service a message (i.e. from the time the message is received by FIST DMD to the time the response is first transmitted)

3) Time from first transmission of service message until acknowledgement (ACK) is received for that message

4) Frequency count by try number of messages acknowledged

5) Frequency count by try number of messages never acknowledged

6) Number of fire missions completed/number of fire missions initiated

7) Number of fire missions completed/number of fire missions expected (i.e. number of fire missions in the data base)

C. Scope

The fire support team will be a four-man team consisting of :

1) the fire support team chief

2) the fire support sergeant

3) two radio telephone operator/drivers.

The FIST chief will be available to the FIST HQ for initial supervision only. As per typical operating procedures, the FIST chief may be absent for extended periods of time (hypothetically accompanying the company commander).

The FIST HQ will be task-loaded by software interactively simulating three platoon-level forward observers. The software FOSCE (Forward Observer SCEnario) will use tactical scenarios developed by Mr. Arthur Long of the US Army Field Artillery Board. This scenario or input database is detailed in the Section D, "Input Data Base" under RESOURCE REQUIREMENTS.

The FIST HQ will have direct access to fire support from a company-level mortar platoon fire direction center (FDC) and a generic field artillery fire direction center. All FDC operations will be simulated interactively by software. The FIST HQ will determine the proper action (based on the FIST chief's guidance and training) for each fire request; either to deny the request, service the request with mortars or forward the request. Fire support will be unlimited, that is, not constrained by ammunition resupply.

All members of the FIST Headquarters will be trained in the operation of the FIST DMD to give the FIST chief flexibility in managing his team.

D. Limitations

1) The FIST DMD will be operated in the review mode only.

2) After receiving a fire request, the fire mission will be forwarded in the automatic mission mode.

3) No FIST HQ initiated missions will be included.

4) No tactical chores will be performed, e.g., guard duty, close station march order, emplacement, etc.

5) All communication will be digital, no voice communication.

E. Test Configuration

Figure 3 shows the nodes that will be played in the first military player test. The FIST HQ equipped with the FIST DMD in the mock-up vehicle interacting through ETHER, the intracomputer communications network, with three forward observer scenario programs, the mortar fire-direction simulator and battalion fire-direction simulator. Figure 4 shows how these players communicate together and the net assignments.

IV. RESOURCE REQUIREMENTS.

A. Software

ACE software permits real-time fire support command and control functions to be exercised in a controlled laboratory environment. The software is written in the C programming language and is designed to run under the 4.1bsd (Berkeley) UNIX operating system. The major components of the ACE software are described below.

1. ETHER

ETHER is a single program which functions as an intra-computer communications network. Computer ports are assigned to communication nets. ETHER accepts a message from a port and transmits it to all other ports on the assigned net. Message collisions are prevented by separately buffering each message within ETHER.

Each net is assigned a probability of message loss which ranges from zero to one. If the probability of message loss is zero, the net is an ideal net and all messages are sent to each port on the net. If the probability of message loss is greater than zero, a uniform random number generator is used to decide whether or not a message is lost. Lost messages are not transmitted to any port on the net.

ETHER maintains a log file of each message which it receives. In addition to the raw message, the log contains the times (Julian day, hour, minute, second) for the start of the message, the end of the preamble and the end of the message.

2. Ace Display (ADIS)

ADIS utilizes a CRT (cathode ray tube) terminal to display in real time the messages being transmitted through ETHER. The terminal screen is divided into eight columns which are labeled for the players (see Figure 5). Each message is displayed as two lines in both the senders and receivers columns. The message first appears in the senders column. The first line contains the message type and target number if it has been assigned. The second character in the second line is a "^", indicating "sender" and the time sent is given. The message will then appear in the "receivers" column. The first line is the same as in the "sender's" , the second character in the second line gives the address of the "sender" and the time received is displayed. When the acknowledgement is sent by the "receiver" an "*" is displayed as the first character in the second line of the "receiver" and when the acknowledgement is received by the "sender" an "*" is displayed as the first character in the second line of the "sender". If the message is degraded by ETHER "MSG LOST" appears in the receivers column. Below the columns, the last message sent is interpreted. At the bottom of the screen, the time from the start of that run is displayed.

3. Forward Observer Scenario (FOSCE)

Forward observer scenario program reads a database of forward observer (FO) messages and transmits the messages as if they were being generated by a real FO. Each message is time tagged so that FOSCE will know when to send it. In addition, FOSCE will retransmit a message if it does not receive an acknowledgement and will wait for message-to-observer (MTO) and SHOT messages before transmitting subsequent adjust (SA) messages.

4. Fire Direction Simulator (FDS)

The fire direction simulator consists of four programs which perform a limited number of TACFIRE/BCS functions. FDS accepts fire request messages, prioritizes them, assigns target numbers and generates MTO and SHOT messages. The number of simultaneous missions which the FDS will process may be specified. If the number of missions exceeds the maximum, the FDS will process missions based on mission priority.

5. Mortar Fire Direction Simulator (MFDS)

The mortar FDS simulates communication with the 81 mm company mortars. It is a special version of the FDS program which will only accept one fire mission at a time.

6. Bit Box Program (BBP)

The bit box interface program accepts messages from ETHER and transmits them to a computer port which is connected to a bit box. The program also reads messages from the computer port and transmits them to ETHER.

B. Hardware

1. Two Bit Boxes

Bit boxes are microprocessor based devices which enable TACFIRE hardware to communicate with commercial computers. Bit boxes accept TACFIRE messages from wire line or radio, perform error correction and convert the messages to RS232 ASCII characters which commercial computers can accept. They will also accept a message from the computer, add the error correction bits, time disperse the message and transmit it over wire line or radio in TACFIRE format (FSK).

2. FIST DMD

The FIST digital message device that will be used in the experiment is one of four experimental design models (EDM) that are in existence. It is a prototype model, and not a production model.

3. VAX 11/750 computer

The VAX 11/750 computer will be dedicated to running the experiment and will have no other processes running during the test. The operating system is the 4.1bsd (Berkley) UNIX.

C. Training

Test participants will be trained in the operation of the FIST DMD by an instructor from the Gunnery Department of the US Army Field Artillery School at the Human Engineering Laboratory. The Human Engineering Laboratory will provide training equipment for the students. The test participants are trained Fire Support Teams.

D. Input Data Base

The tactical scenario data base contains all fire support control messages for a limited scenario of a mechanized infantry battalion of an armored division. The SCORES, Europe III, Sequence 2A was used to generate targets expected to be fired by a battalion in sustained combat operation. The battalion is constrained by ammunition resupply under normal operations, however, it was decided that ammunition resupply should not be a limiting condition in this test. The entire scenario is played in retrograde mode.

The data base consists of 36 two hour cells of messages, 12 two hour cells of low intensity, 12 two hour cells of medium intensity and 12 two hour cells of high intensity. Intensity is defined by the number of initiating messages per two hour cell in Figure 6 and the message stream that follows each initiating message is given in Figure 7. It can be seen that intensity is a function of the number of initiating messages and their subsequent messages. The 36 two hour cells of data are arranged such that all permutations of the three intensities (L-M-H) appear twice. Ninety percent of the fire missions will have normal priority and the other ten percent urgent priority.

V. DATA COLLECTION.

A. Experimental Design

1. Factors.

The two factors that will be tested in this experiment are message intensity and communications reliability. Three levels of message intensity will be tested with each of three levels of communication reliability giving nine test combinations. Degradation of messages is total. That is, 15% degradation indicates 15% of the messages are lost in their entirety. The levels of each factor will be defined as follows:

Message Intensity

L = low

M = medium

H = high

Communications Reliability

0 = 0% degradation

1 = 15% degradation

2 = 30% degradation

2. Assumptions.

For measures of performance 1,2,3,6 and 7 it is assumed that these measures are independent and normally distributed random variables and that all observations of a given MOP have the same variance. Only the assumption of independent random variables is made for the remaining measures of performance.

3. Design Matrix.

Since the testing of all nine treatment combinations require a minimum of 18 hours of testing and realistically could not be completed in one day, a randomized incomplete block design was constructed as shown in Figure 8 so that the day-to-day variability would not influence our results. The nine treatment combinations were divided into blocks of three and the three blocks were run over a three day period in a random order. The experiment will be repeated for each of the four FISTs so that an unbiased estimate of error can be obtained. In addition, a comparison of the performance of the four FIST HQs can be made since each team is tested under all possible treatment combinations twice.

The assignment of the treatment combinations into blocks was based on a confounding scheme. This scheme assures that the effects of message intensity (I) and communication degradation (C) and the interaction of these two factors (I x C) on a FIST HQ's ability to perform fire-support coordination can be measured. Because time constraints permit only two replications, part of the precision of the estimate of the interaction was sacrificed (i.e. blocks within replicate 1 are confounded with the linear component of the I x C interaction and blocks within replicate 2 are confounded with the quadratic component of the I x C interaction). Randomization of treatment combinations within blocks and blocks within days has been performed, hence, the test will be run in the order shown in the design matrix, Figure 8.

B. Questionaires

Questionaires will be administered at the end of the FIST DMD training and at the end of the test. The questionaires were developed by Mr. Leonard Cunningham and Major Grim of the Field Artillery Board for the FIST DMD Force Development Testing and Experimentation (FDTE).

VI. DATA ANALYSIS.

A. Statistical Analysis

The data analysis will be based on a incomplete block design. Analysis of variance or a two-way classification table analysis, frequently called a contingency table analysis, will be the methods of analyses performed. The null hypotheses to be tested are:

1) There is not a statistically significant difference between the message intensity levels as measured by the stated measures of performance.

2) There is not a statistically significant difference between the degradation levels as measured by the stated measures of performance.

3) There is not a statistically significant interaction effect between message intensity and communication degradation.

An analysis of variance (ANOVA) will be used to test the above hypotheses for MOPs 1,2,3,6 and 7. However, since this analysis is sensitive to departures from the assumption of equal variances, a check will be made to assure that this assumption is valid. An appropriate transformation will be performed on those MOPs that depart from this assumption to assure the validity of the analysis. The analysis of variance table for this design is presented in Figure 9. The error term will be obtained by pooling the FIST, replicate, and day interactions after a check has been made to assure that these effects are not significant.

A contingency table analysis will be developed for MOPs 4 and 5. The chi square statistic will be used to test the above hypotheses.

B. Subjective Analysis

Questionaires will be summarized to provide feedback on the experiment, the CPX Research Facility, training and equipment.

BRIGADE-AREA TACFIRE SYSTEM

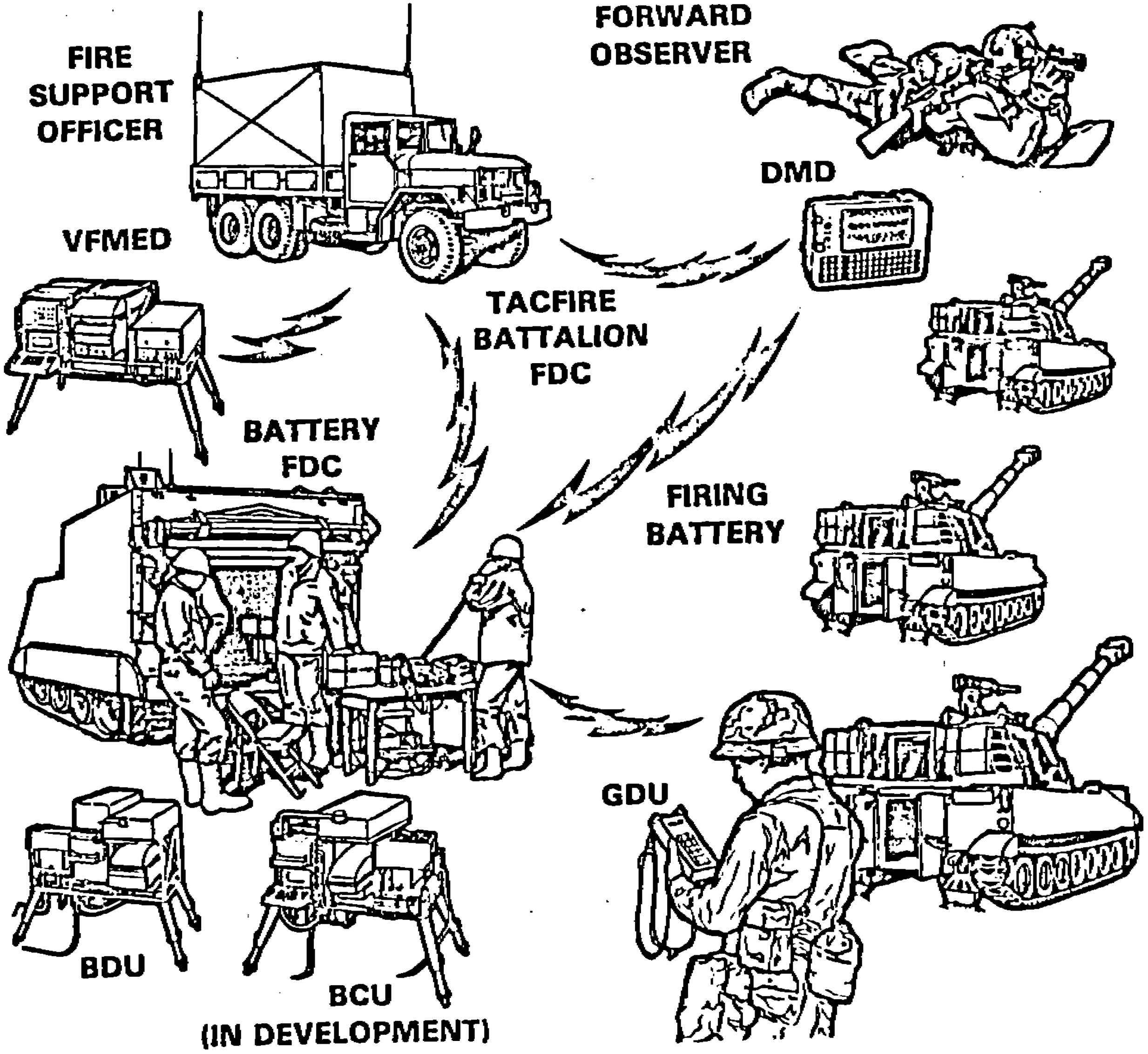


FIGURE 1

COMMAND POST EXERCISE RESEARCH FACILITY

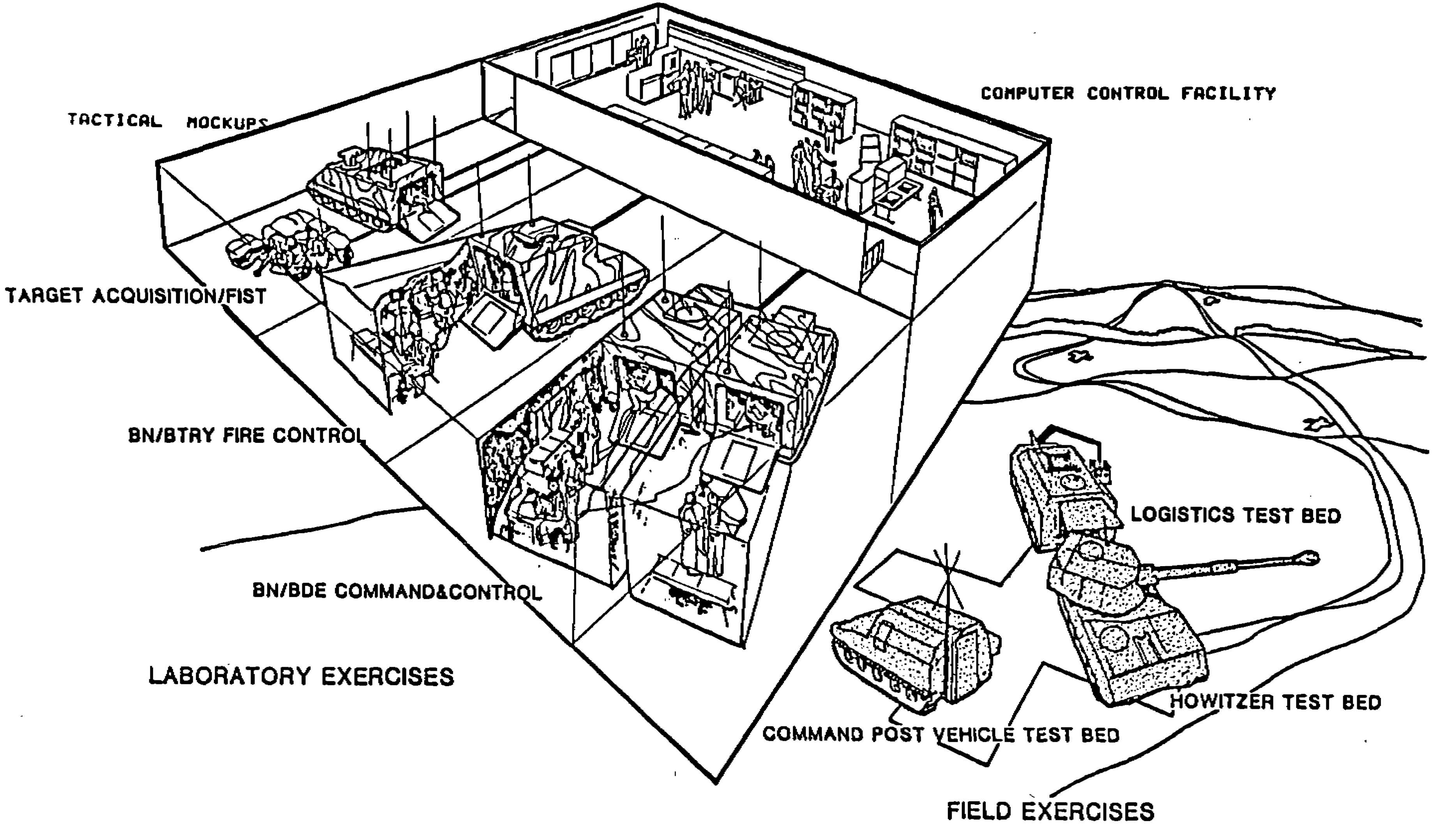
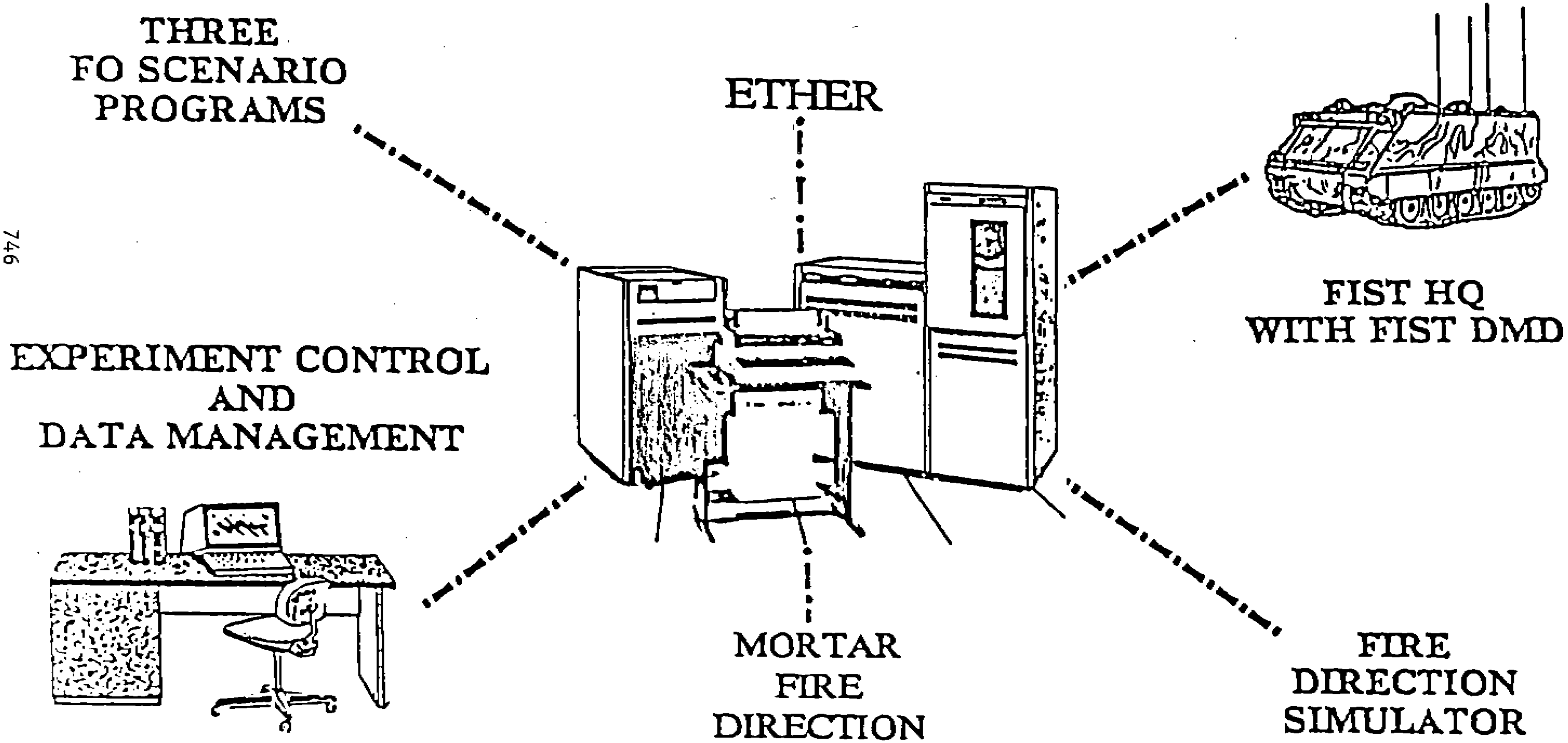


FIGURE 2

FIST HEADQUARTER EXPERIMENT



THREE
FO SCENARIO
PROGRAMS

ETHER

FIST HQ
WITH FIST DMD

EXPERIMENT CONTROL
AND
DATA MANAGEMENT

MORTAR
FIRE
DIRECTION
SIMULATOR
PROGRAM

FIRE
DIRECTION
SIMULATOR
PROGRAM

FIGURE 3

ETHER: Intra-computer communications network
FOSCE: Forward Observer SCEnario program
FDS: Fire Direction Simulator program
MFDS: Mortar Fire Direction Simulator program

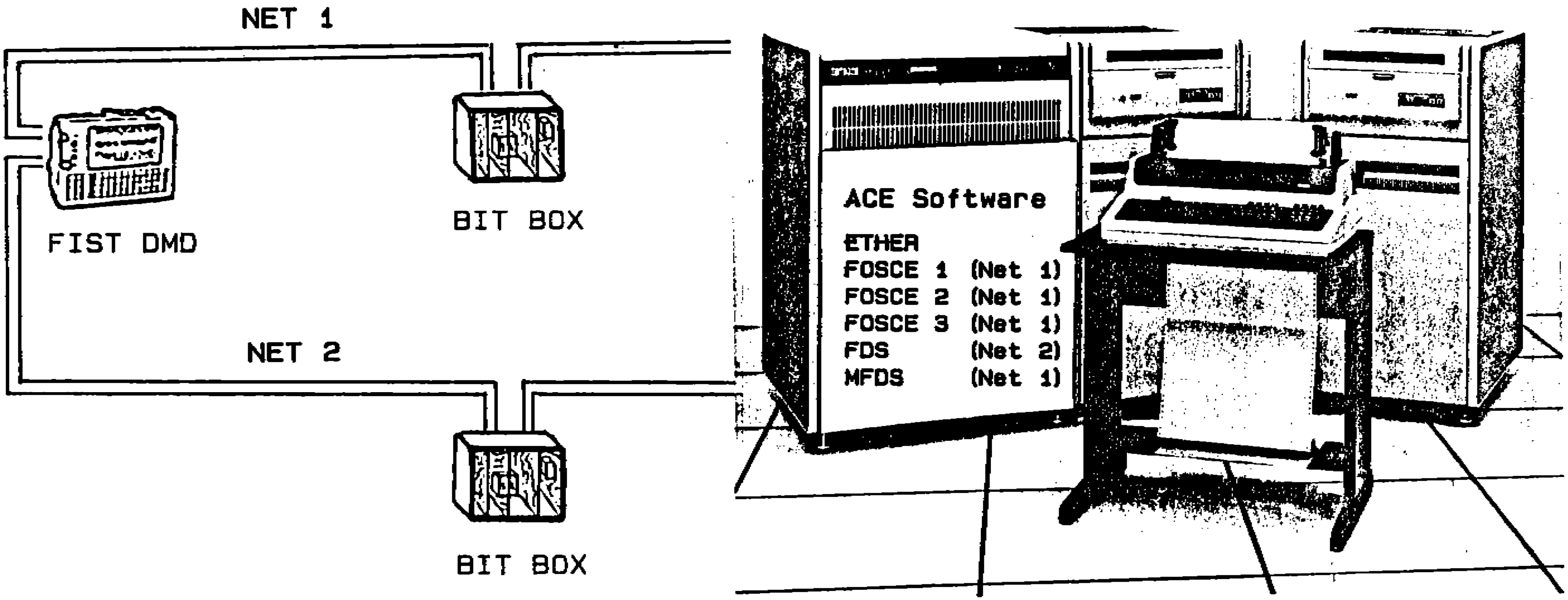
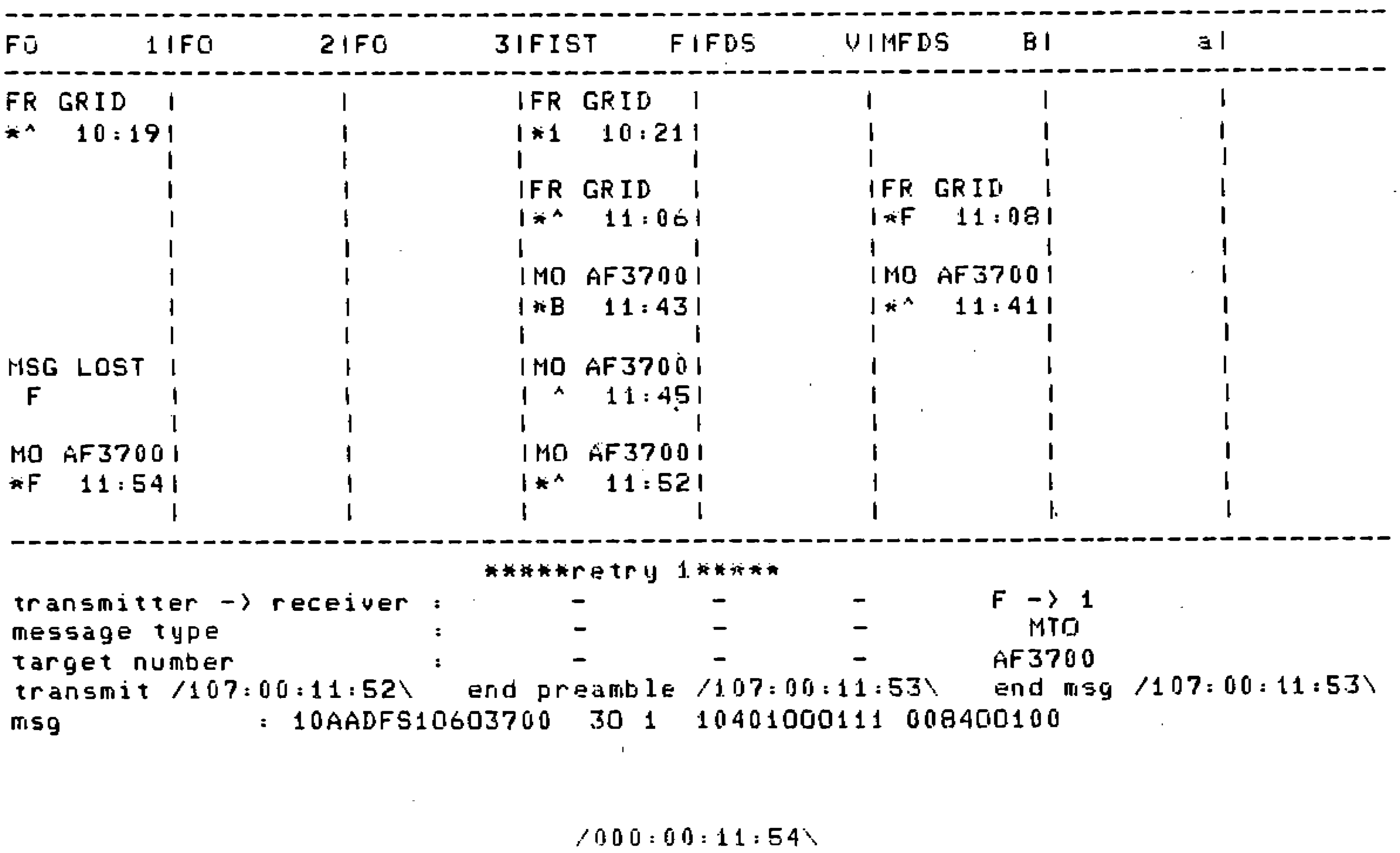


FIGURE 4. Test Configuration



```
--------------------------------------------------------------------------------
FO      1|FO      2|FO      3|FIST    F|FDS     V|MFDS    B|        a|
--------------------------------------------------------------------------------
FR GRID  |         |         |FR GRID  |         |         |         |
*^  10:19|         |         |*1  10:21|         |         |         |
         |         |         |         |         |         |         |
         |         |         |FR GRID  |         |FR GRID  |         |
         |         |         |*^  11:06|         |*F  11:08|         |
         |         |         |         |         |         |         |
         |         |         |MO AF3700|         |MO AF3700|         |
         |         |         |*B  11:43|         |*^  11:41|         |
         |         |         |         |         |         |         |
MSG LOST |         |         |MO AF3700|         |         |         |
 F       |         |         | ^  11:45|         |         |         |
         |         |         |         |         |         |         |
MO AF3700|         |         |MO AF3700|         |         |         |
*F  11:54|         |         |*^  11:52|         |         |         |
         |         |         |         |         |         |         |
--------------------------------------------------------------------------------
                         *****retry 1*****
transmitter -> receiver :      -       -       -       F -> 1
message type            :      -       -       -         MTO
target number           :      -       -       -       AF3700
transmit /107:00:11:52\   end preamble /107:00:11:53\   end msg /107:00:11:53\
msg            : 10AADFS10603700  30 1  10401000111 008400100


                                /000:00:11:54\
```

FIGURE 5. Sample ACE Display (ADIS)

FACTORS

1) INTENSITY (per two hour block)

MESSAGE TYPE	LEVELS		
	Low	Medium	High
Fire Mission 1, Fire For Effect	4	8	12
Fire Mission 2, Adjust Fire	2	4	6
Fire Mission 3, Immediate Smoke	0	1	1
Artillery Target Intelligence	18	12	6

2) COMMUNICATION DEGRADATION

00% Message Loss
15% Message Loss
30% Message Loss

FIGURE 6

INTENSITY				
		LEVELS		
MESSAGE SEQUENCE.		L	M	H
1) Artillery Target Intelligence ATI	 FO→FIST→FDC	18	12	6

2) Fire Mission, Fire for Effect:
FR GRID
MTO
SHOT
EOM
	
	FO→FIST→FDS
	FO←FIST←FDS
	FO←FIST← FDS
	FO→FIST→FDS
		4	8	12

3) Fire Mission, Adjust Fire FR GRID MTO SHOT SA(1) SHOT SA(2) SHOT SA(3) SHOT EOM	 FO→FIST→FDS FO←FIST←FDS FO←FIST←FDS FO→FIST→FDS FO←FIST←FDS FO→FIST→FDS FO←FIST←FDS FO→FIST→FDS FO←FIST←FDS FO→FIST→FDS	2	4	6
4) Fire Mission, Immed. Smoke Same as Adjust Fire Mission		0	1	1

FIGURE 7

DESIGN MATRIX						
FIST	REP1			REP2		
TEAM	DAY1	DAY2	DAY3	DAY4	DAY5	DAY6
TEAM ONE	L2 M1 H0	M0 H2 L1	L0 H1 M2	M0 H1 L2	L0 M1 H2	H0 L1 M2
TEAM TWO	H1 L0 M2	H2 M0 L1	L2 H0 M1	M1 L0 H2	M2 H0 L1	H1 M0 L2
TEAM THREE	M2 H1 L0	H2 M0 L1	M1 L2 H0	L0 M1 H2	H1 M0 L2	M2 H0 L1
TEAM FOUR	H2 M0 L1	M1 L2 H0	M2 H1 L0	H1 M0 L2	L1 H0 M2	L0 M1 H2

INTENSITY	COMMUNICATION DEGRADATION
L= LOW	0= 00% DEGRADATICN
M= MEDIUM	1= 15% DEGRADATICN
H= HIGH	2= 30% DEGRADATION

FIGURE 8

ANALYSIS OF VARIANCE (ANOVA)	
SOURCE	DEGREES OF FREEDOM
FIST	3
Rep	1
Block within Rep	4
Intensity	2
Degradation	2
Intensity x Degradation	4
Pooled error	55
Total	71

FIGURE 9

MOVING FINITE ELEMENT RESEARCH FOR SHOCK HYDRODYNAMICS, CONTINUUM MECHANICS AND COMBUSTION

Robert J. Gelinas
Said K. Doss
Neil N. Carlson
Science Applications, Inc.
1811 Santa Rita Road, Suite 104
Pleasanton, California 94566

ABSTRACT. The overall objective of this research is to investigate the numerical properties and structure of the moving finite element (MFE) method in order to reduce it to practice for the numerical solution of important PDE systems. This research focusses upon mathematical and computational properties of transient MFE solutions in 1-D and 2-D of (i) the full viscous, compressible Navier-Stokes equations for shocks and possibly for combustion processes in gases, and (ii) the continuum equations for impacts of initially solid bodies, where constitutive models include elastic, plastic, and visco plastic effects. In this work, primary attention is devoted to the distinction and exacting resolution of actual physical dissipation effects (over highly disparate scales) vis a vis numerical dissipation effects which frequently obscure the actual physical dissipation processes in PDE solutions of fluid dynamics equations. Test cases which demonstrate these distinctions are presented, and those factors which are major determinates of grid node optimality in the MFE method and in certain other adaptive solution methods for PDE's are discussed.

INTRODUCTION. The moving finite element (MFE) method is a promising new approach for solving numerically the partial differential equations (PDE's) of hydrodynamics, continuum mechanics, combustion, and other transport equation systems. In the MFE method, grid coordinates themselves are dependent variables which are calculated continuously at each time step in order to minimize PDE residuals. This feature has successfully suppressed numerical dissipation to very low levels which has resulted in the accurate resolution in 1-D shocks of physical transport and dissipation effects which are contained in the full Navier-Stokes equations. It has been shown in recent work that: (i) the MFE method does extend logically and practically to 2-D and (ii) extensive research in several new areas must be carried out in order to realize the full potential of the MFE method.

This paper reviews progress in 1-D and 2-D MFE research and indicates areas where applied mathematics research will continue to advance progress in MFE developments.

1. Steady State Shock Structure Calculations. Until recently, neither computer hardware nor numerical solution methods for partial differential equations (PDE's) could realistically be expected to resolve the highly disparate physical scales of physical dissipation processes in shocked fluids. However, the MFE method has recently shown promise in 1-D of resolving shock structures according to the actual physical dissipation processes and thermal conduction which occur in nature. The resolution of such highly disparate scales is especially important in transient systems where shocks, contact surfaces, and other fluid structures may interact. An essential benchmark is a verification that the MFE method does, in fact, reproduce those shock

structures for which analytic solutions are known. Once such example is the steady state solution of the full Navier-Stokes equations for a freely propagating shock in an ideal gas. Recall that the Navier-Stokes equations are derived as the first approximation of kinetic theory which is outlined for 1-D systems in Equations (1)-(9) below. From kinetic theory,[1]

$$\frac{\partial}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0 \tag{1}$$

$$\frac{\partial(\rho v)}{\partial t} + \frac{\partial}{\partial x}(\rho v^2) = -\frac{\partial}{\partial x}\left\{p^{(0)} + p^{(1)} + p^{(2)} + \ldots\right\} \tag{2}$$

$$\frac{\partial E}{\partial t} + \frac{\partial}{\partial x}(Ev) = -\frac{\partial}{\partial x}(pv) - \frac{\partial}{\partial x}\left\{q^{(0)} + q^{(1)} + q^{(2)} + \ldots\right\} \tag{3}$$

A zero-th approximation of the kinetic theory for gases uses the constitutive relations for an inviscid, non-conducting fluid; i.e.,

$$p = p^{(0)} = \frac{\rho RT}{A} = (\gamma - 1)(E - \tfrac{1}{2}\rho v^2) \tag{4}$$

$$q = q^{(0)} = 0 \tag{5}$$

These relationships yield the well-known Euler equations.

A first approximation of the kinetic theory for gases gives the Navier-Stokes equations according to:

$$p = p^{(0)} + p^{(1)} = (\gamma - 1)(E - \tfrac{1}{2}\rho v^2) + \tfrac{4}{3}\mu\frac{\partial v}{\partial x} \tag{6}$$

$$q = q^{(0)} + q^{(1)} = -\kappa\frac{\partial T}{\partial x} \ , \tag{7}$$

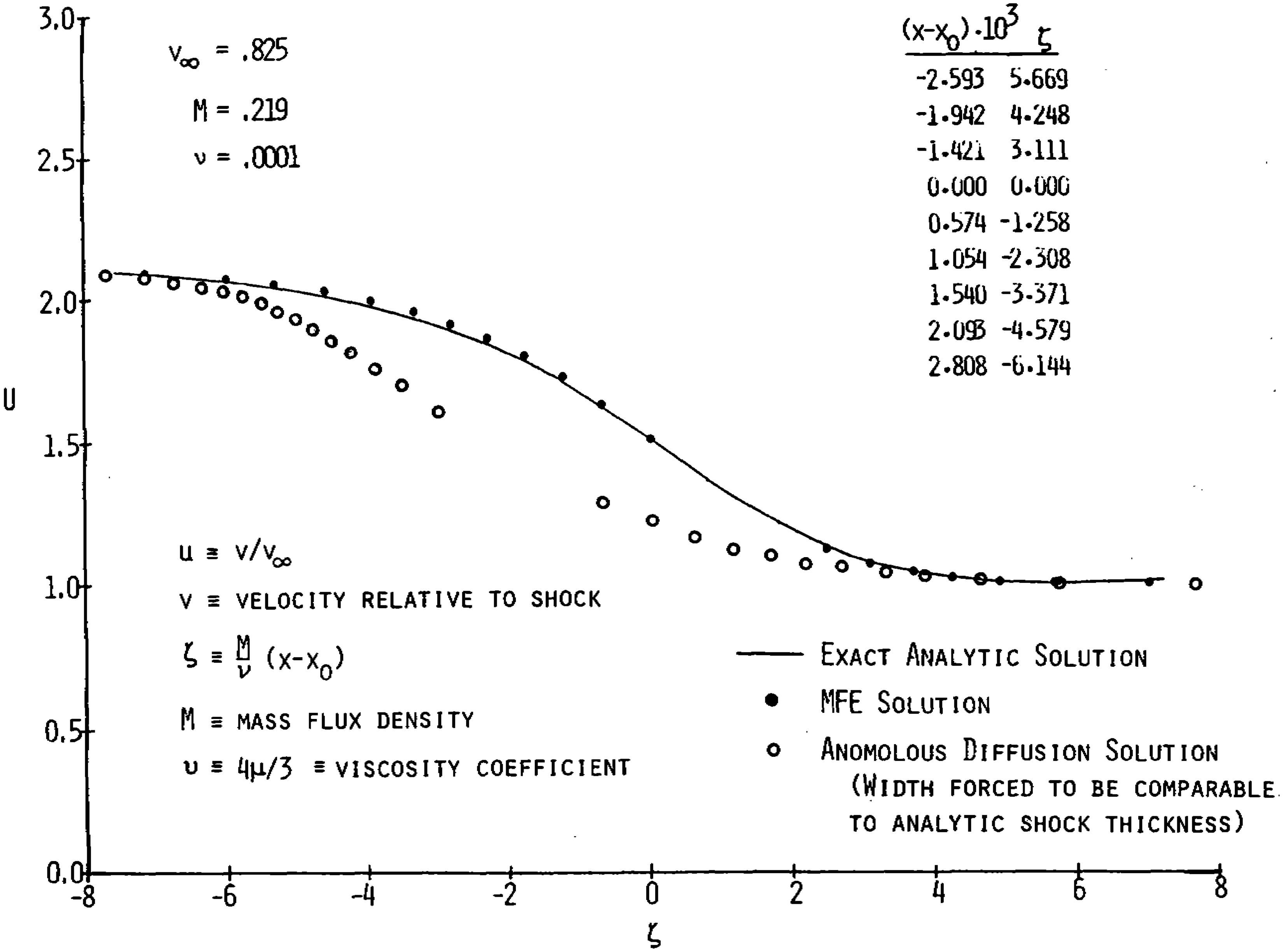
and a second approximation of kinetic theory gives the Burnett equations,

$$p = p^{(0)} + p^{(1)} + p^{(2)} \tag{8}$$

$$q = q^{(0)} + q^{(1)} + q^{(2)} \quad , \text{etc.} \tag{9}$$

The analytic solutions of the steady state Navier-Stokes equations are readily obtained and can be found in Reference 2. Using these analytic results, we can consider comparisons between the MFE solution, the exact analytic solution, and an anomolous numerical diffusion solution for the shock test problem which has been considered previously by Sod.[3] Using the notation $m \equiv \rho v$, the initial conditions are:

$$\begin{array}{ll}
m(x,0) = 0 & 0 \leq x \leq 1.0 \\
\rho(x,0) = 1 & 0 \leq x \leq 0.500 \\
E(x,0) = 2.5 & \\
\rho(x,0) = p(x,0) = \text{linear} & 0.500 \leq x \leq 0.501 \\
E(x,0) = \text{linear} & \\
\rho(x,0) = 0.125 & 0.501 \leq x \leq 1.0 \\
E(x,0) = 0.25 \quad , &
\end{array} \tag{10}$$



Relative Fluid Velocity Versus Position

Figure 1. MFE solution of Navier-Stokes equation for G. Sod's test problem.

and Dirichlet boundary conditions are used at x = 0 and x = 1.0.

Figure 1 presents comparative results for the shape of the shock front in normalized units during free, steady state propagation at t = 0.15. It is seen that the MFE solution of the Navier-Stokes equations (solid dots), which used only 19 grid nodes, compare very well with the analytic solution (solid line). As a numerical experiment, one can ask: How sensitive are the solutions of the this problem to some anomolous or artificial dissipation process? To address this question, we replaced the correct physical viscous dissipation terms in the Navier-Stokes equations by a linear dissipation operator (essentially a linear Q model, which is frequently used in order to stabilize some other PDE solution methods, with the hope that the physical solutions of the true Navier-Stokes equations would not be altered too greatly by the more more convenient linear dissipation model). This numerical experiment was also performed with the linear dissipation term appearing in the density equation, as well in the momentum equation, as a sensitivity test of the possible effects of uncontrolled numerical diffusion which is intrinsic in any PDE method when the grid mesh locations are not optimal. It was found that using the same value of ν = 0.0001 in the uncontrolled, or anomolous, dissipaton model as was used in the true Navier-Stokes calculations, the computed shock shape and width from the anomolous dissipation model were greatly in error. We then asked: Can the value of ν in the anomolous dissipation model be selected somehow arbitrarily so as to achieve agreement between the Navier-Stokes solutions and the linear dissipation model solutions? It was found that the width of the shock calculated by the anomolous dissipation model could be made to approximate the width of the shock in the Navier-Stokes solutions by increasing the value of ν by a factor of five in this specific case. But it was not possible to reproduce the correct Navier-Stokes shock shape, as can be seen from the open circles in Figure 1, by any of our attempted adjustments of ν in the anomolous dissipation model.

It is, of course, recognized that, in many practical examples, it may not be necessary to know the correct shock structure. We therefore extended this initial investigation of steady state shock effects in order to determine if, when, and how shock structure effects may become important in practical applications. A first extension continued the present MFE calculations of this test example through many shock reflections. These results have been reported in Reference 4. It was found that the MFE solutions of the Navier-Stokes equations can accurately resolve the repeated mutual interactions and reflections of shocks, contact surfaces and rarefactions and that this problem can be solved all the way to its final equilibrium state. It is clearly evident in these transient results that the accurate resolution of the actual physical dissipation processes over highly disparate scales in the Navier-Stokes equations is essential for practical applications in which shocks are not simply in a state of free propagation. That is, the strength of interactions between fluid structures and thus the time required to reach equilibrium are sensitive to both the magnitude and operational descriptions of these physical dissipation processes. For example, model thermal conductivities which are larger than their appropriate physical magnitudes would have the effect of broadening contact discontinuities too much, too soon; and ensuing interactions between shocks and such broadened contact discontinuities would drive the system to equilibrium too soon. Conversely, model thermal conductivities which are too small maintain contact discontinuity gradients at values which are larger than their appropriate physical magnitudes until late times, and the corresponding evolution toward equilibrium would then be slower than the appropriate physical evolution. These features can be seen in the results for both this moderate shock example and for the well-known problem of anomalous wall-heating which is considered next.

2. The Anomolous Wall-Heating Problem. Anomolously high temperatures frequently occur in computations of shock reflections from an infinitely reflecting wall in slab geometry or from the reflection of an imploding shock at the origin in cylindrical or spherical geometries. The MFE results which follow indicate the anomolous aspects are eliminated when the physical dissipation processes in the Navier-Stokes equations are accurately resolved in the transient reflection process. We consider the following test problem in 1-D slab geometry:

Initial Conditions:

$$\begin{array}{ll}
\rho(x, 0) = 1 & 0. \le x \le 2. \\
p(x, 0) = \varepsilon(x, 0) = 0 & 0. \le x \le 2.0 \\
v(x, 0) = -1 & \Delta x_0 \le x \le 2.0 \\
v(x, 0) = \text{linear} & 0 \le x \le \Delta x_0 \\
v(x, 0) = 0 & x = 0. \\
\gamma = 5/3 &
\end{array}$$

Boundary Conditions:

Reflection at $x = 0.$
Dirichlet at initial values at $x = 2.0$

Rankine-Hugoniot Solutions for Infinite Shock ($t \to t_\infty$):

$$\begin{array}{lll}
s = 1/3 & & \\
\rho^+ = 4.0 & ; & \rho^- = 1.0 \\
\varepsilon^+ = 0.5 & ; & \varepsilon^- = 0. \\
v^+ = 0. & ; & v^- = -1. \\
p^+ = 1.33 & ; & p^- = 0.
\end{array}$$

The time evolution of this shock was solved by the MFE method in two ways: First the full Navier-Stokes equations were solved with alternative values of $\nu = 4\mu/3$ and κ. These solutions are denoted by N-S in the accompanying figures. In one set of N-S solutions, $\nu = \kappa = 0.01$, which is unrealistically large but which permits comparisons to other fixed node PDE solutions that may use on the order of 100 to several hundred grid nodes. In another set of N-S solutions, $\nu = \kappa = 0.001$, which approximates physically realistic values for actual dissipation processes in gases. Second, the variable ε denotes internal energy per unit mass in the accompanying figures, and anomolous diffusion (denoted by A in the accompanying figures) was simulated by including a diffusion term, $\nu_r \rho_{xx}$, in Equation (1). This effectively simulates some form of uncontrolled numerical diffusion which is present intrinsically in many alternative PDE methods. (Such anomolous diffusion can find its way into alternative PDE methods by numerous and various means).

At $t = 0_+$, the shock incident on the origin is in the incipient state of outward reflection. At $t = 0.05$, Figures 2 and 3 show that the calculations of the reflected shock with uncontrolled diffusion (or simulated numerical diffusion) tend immediately to overheat in ε and to correspondingly undershoot in ρ relative to the Navier-Stokes solutions. Although these

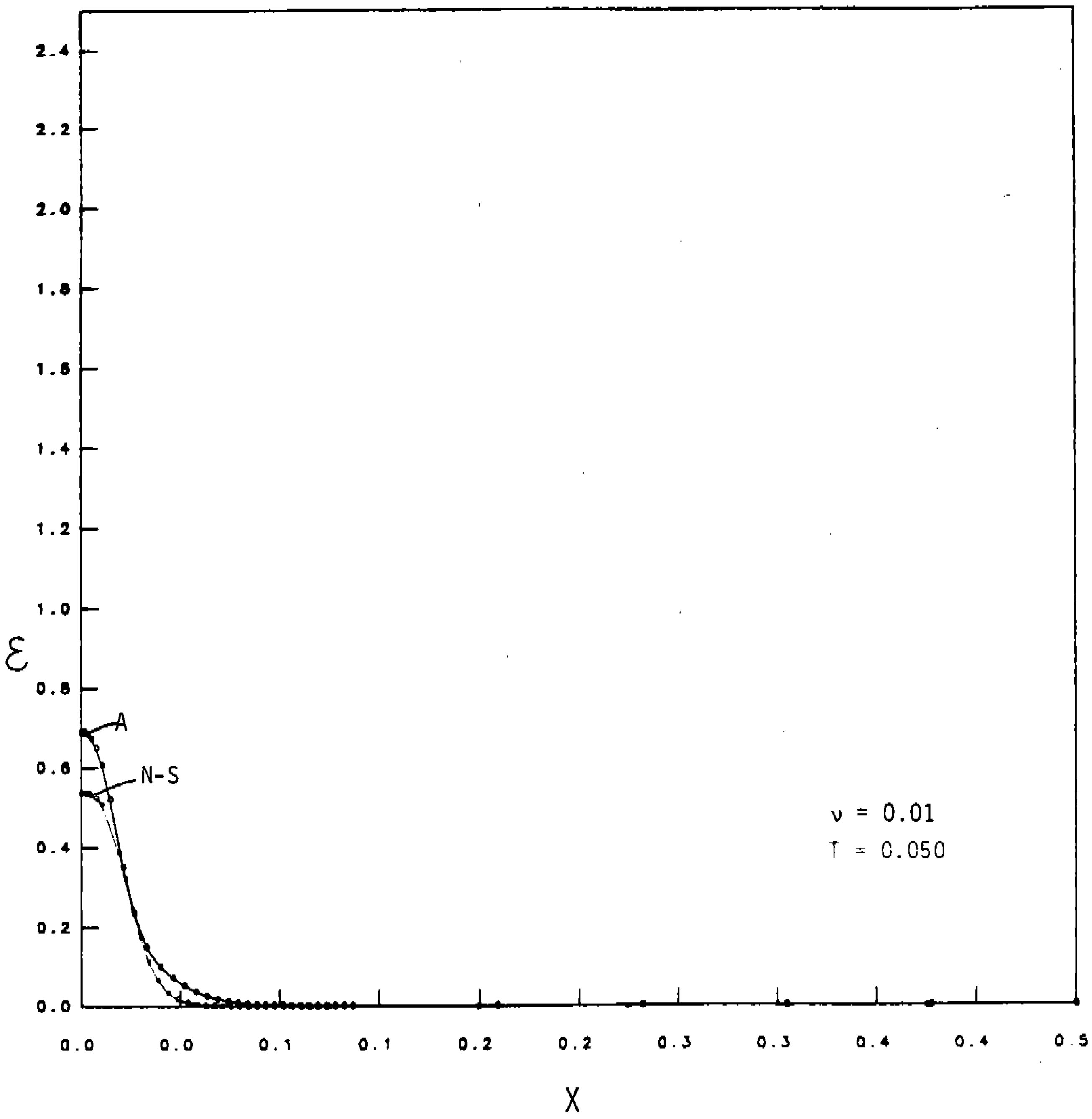


Figure 2. MFE solutions of the Navier-Stokes equations and anomolously dissipative equations.

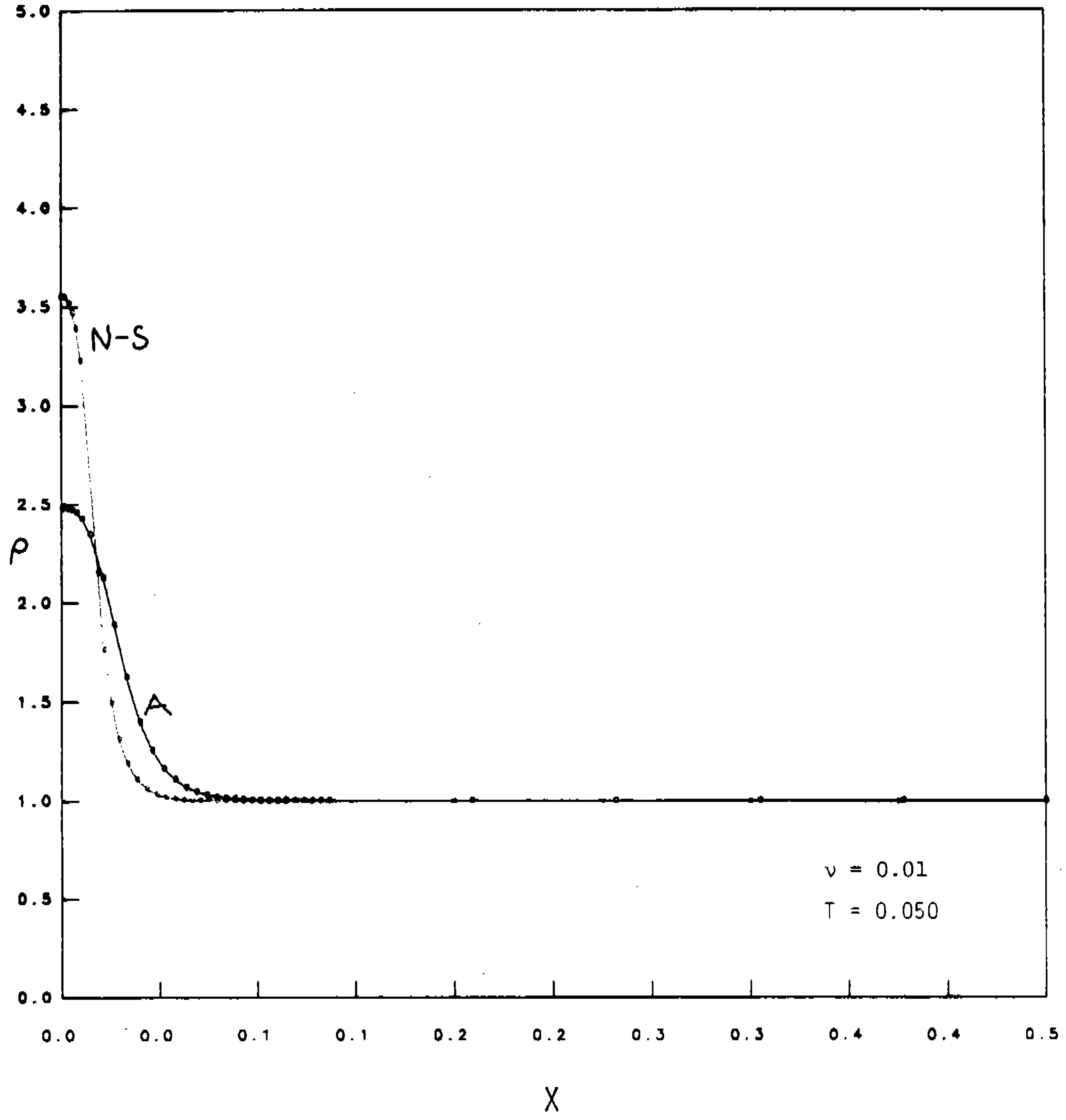


Figure 3. MFE solutions of the Navier-Stokes equations and anomously dissipative equations.

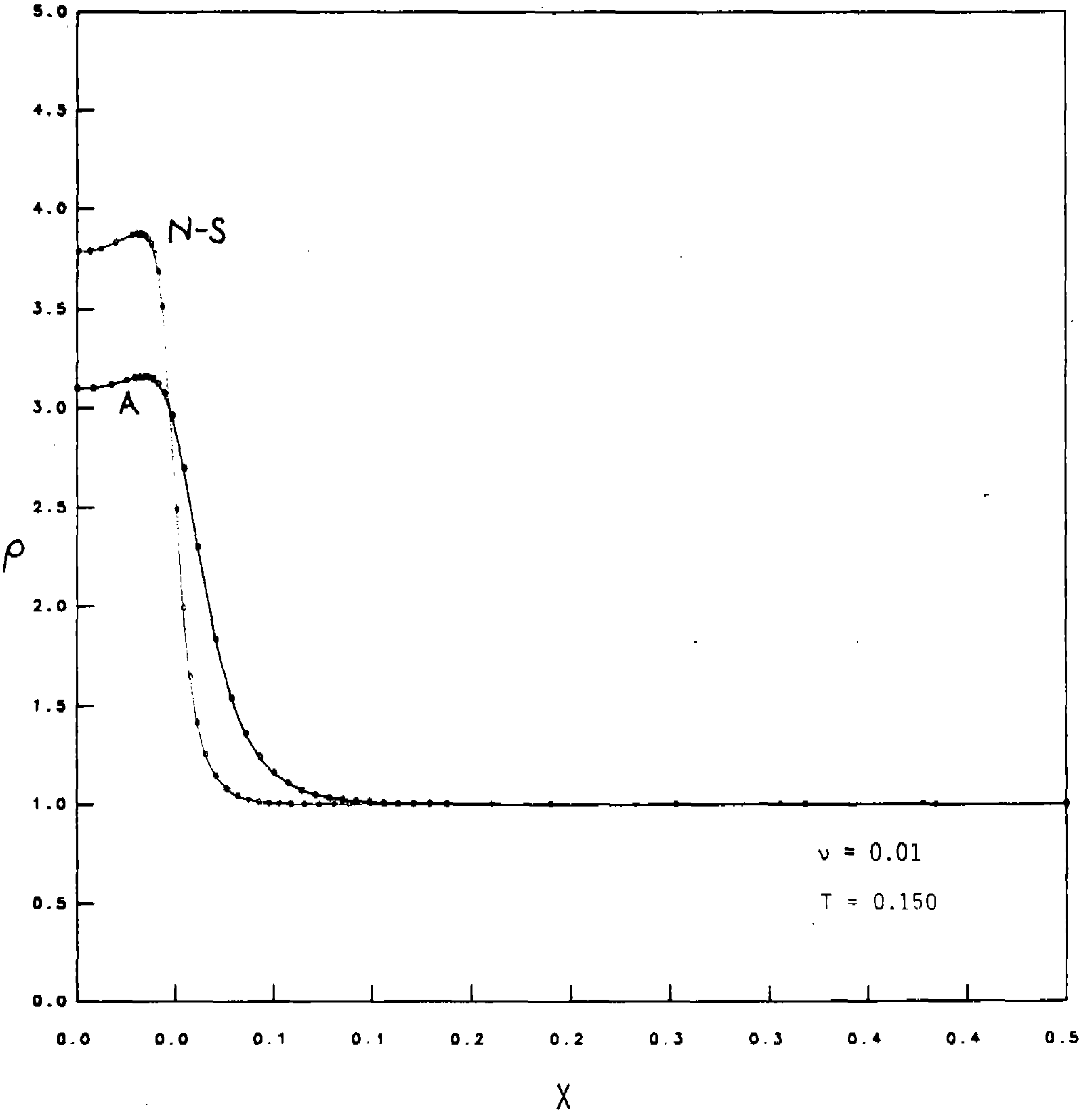


Figure 4. MFE solutions of the Navier-Stokes equations and anomously dissipative equations.

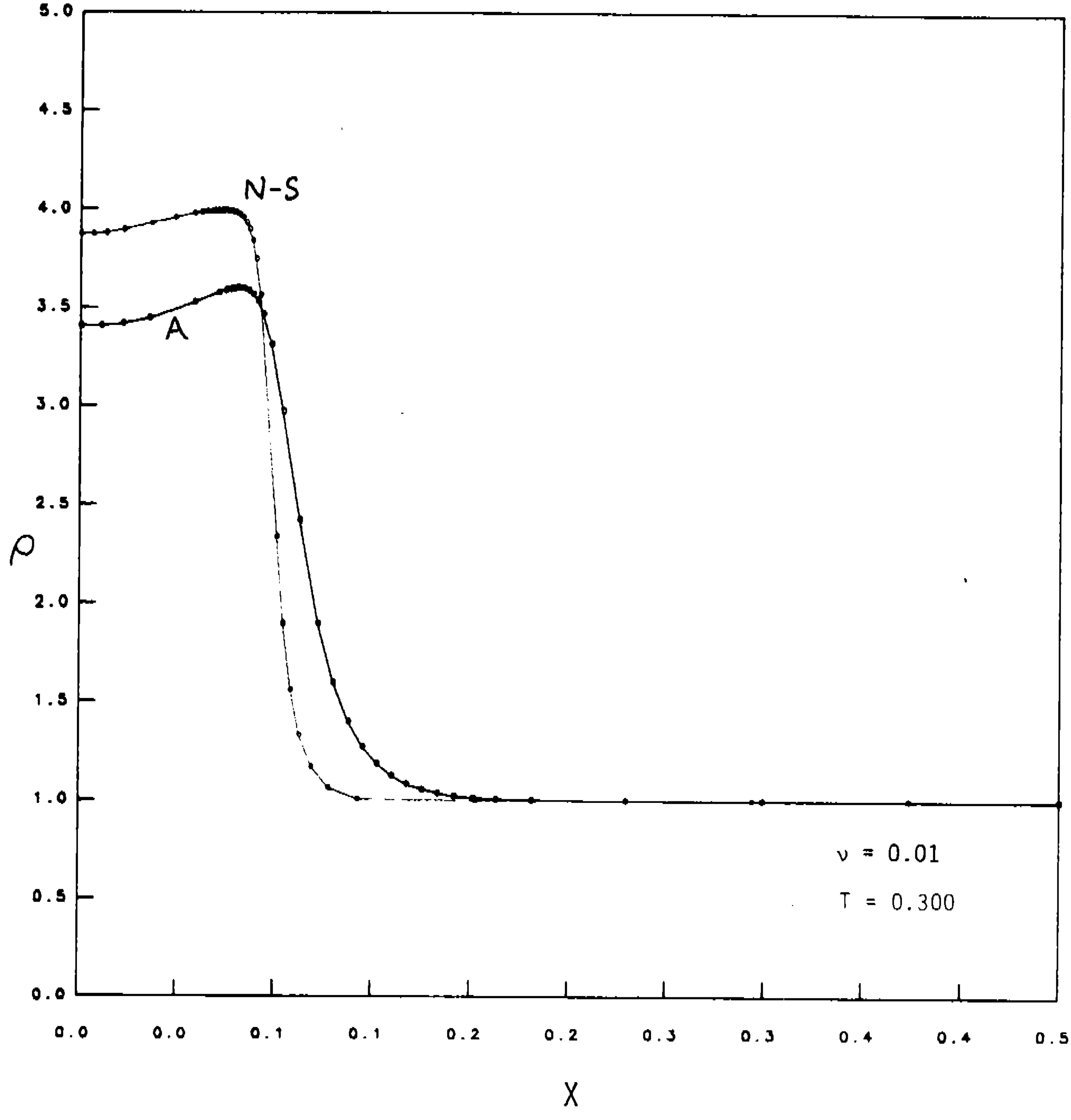


Figure 5. MFE solutions of the Navier-Stokes equations and anomously dissipative equations.

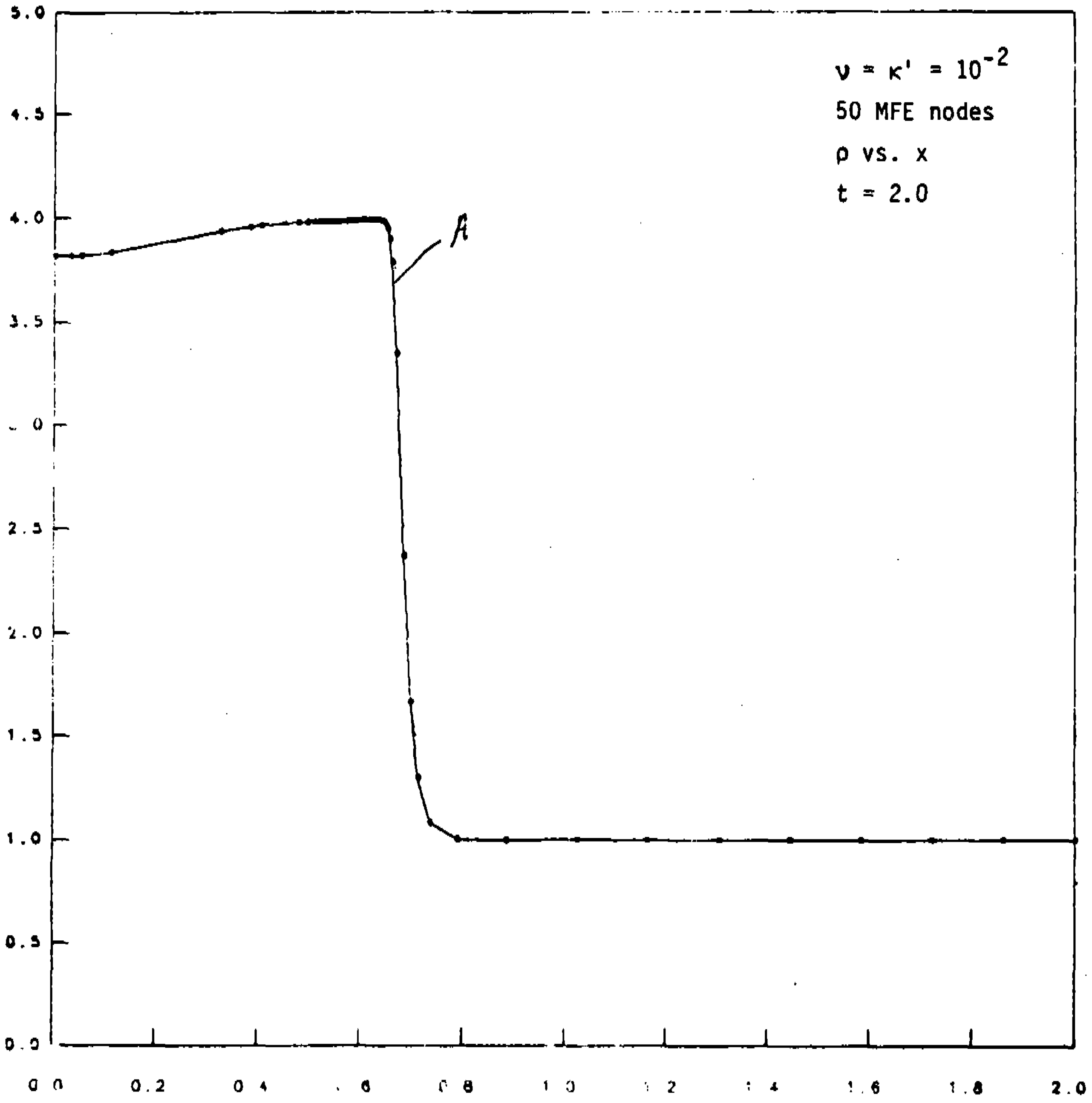


Figure 6. MFE solutions of the Navier-Stokes Equations and anomously dissipative equations.

transient solutions are not near their steady state values at this early time, it will be seen that the ensuing evolution toward equilibrium is quite sensitive to both the magnitudes and the nature of the dissipation processes in the computations.

Figure 4 shows that at t = 0.15, the lip of the shock in the Navier-Stokes solution is approaching the steady state value of $\rho = 4.0$, and the anomolous dissipation solution lags by a significant margin. The fluid buildup at the front of the shock is evident here because the fluid near the origin has stagnated while additional fluid continues to stream in toward the origin from the region to the right of the shock. Figure 5 shows that the anomolous dissipation results continue to lag behind the Navier-Stokes solutions to a significant degree at t = 0.300. At t = 2.0, the Navier-Stokes solutions have approached steady state Rankine-Hugoniot conditions (not shown in Figure 6), and the anomolous dissipation solution has still not reached the Rankine-Hugoniot values in the vicinity of the origin. The anomolous wall heating effects due to uncontrolled dissipation in the density equation have thus persisted to very long times vis à vis the accurate solutions of the Navier-Stokes equations. Non-physical dissipation operators in fluid calculations can be shown to have similar effects.

Figures 7 and 8 present the results of another test of sensitivity of the Navier-Stokes solutions to non-optimal grid locations. In this test problem, a physically realistic value of $\nu = 0.0001$ is used in MFE solutions of the Navier-Stokes equations. We have, however, deliberately constrained the MFE grid nodes in this test case so that they do not migrate to their truly optimal locations, as in the results considered previously. Figure 7 shows several significant features: (i) the shock gradients associated with $\nu = 0.0001$ are extremely large; the accurate resolution of these gradients would require several thousand nodes if a fixed node PDE solution method were to be used, (ii) the Rankine-Hugoniot solutions are approached much more rapidly for the physically realistic values of ν than for the larger values of which are typically used either tacitly or explicitly in many other PDE solution methods, and (iii) the slight constraint on node movements and thus on nodal positions do not show up immediately, but once the perturbation becomes significant (as seen in Figure 8), its effects can grow rapidly. In summary, these results demonstrate that reflected shock solutions can be very sensitive to non-physical dissipation effects and to slight deviations from optical grid node positioning, even in adaptive gridding methods. All of the results in this section were obtained with approximately 30 MFE nodes. As many as 61 MFE nodes were used to verify that the MFE solutions were in fact converged solutions. In the absence of the stringent tests of convergence which were applied here, it can be extremely difficult to discern physical oscillations and dissipation effects from non-physical and/or purely numerical oscillations and dissipation effects.

3. Burger's Equation in 2-D. The PDE's for this skewed Burger's model problem are given by:

$$u_t = -uu_x - vu_y + \nu(u_{xx} + u_{yy}) \qquad (11)$$

$$v_t = -uv_x - vv_y + \nu(v_{xx} + v_{yy}) \qquad (12)$$

where u is the x-component of velocity and v is the y-component, and ν is an effective diffusion coefficient. Shocks are generated with gradient

N-S

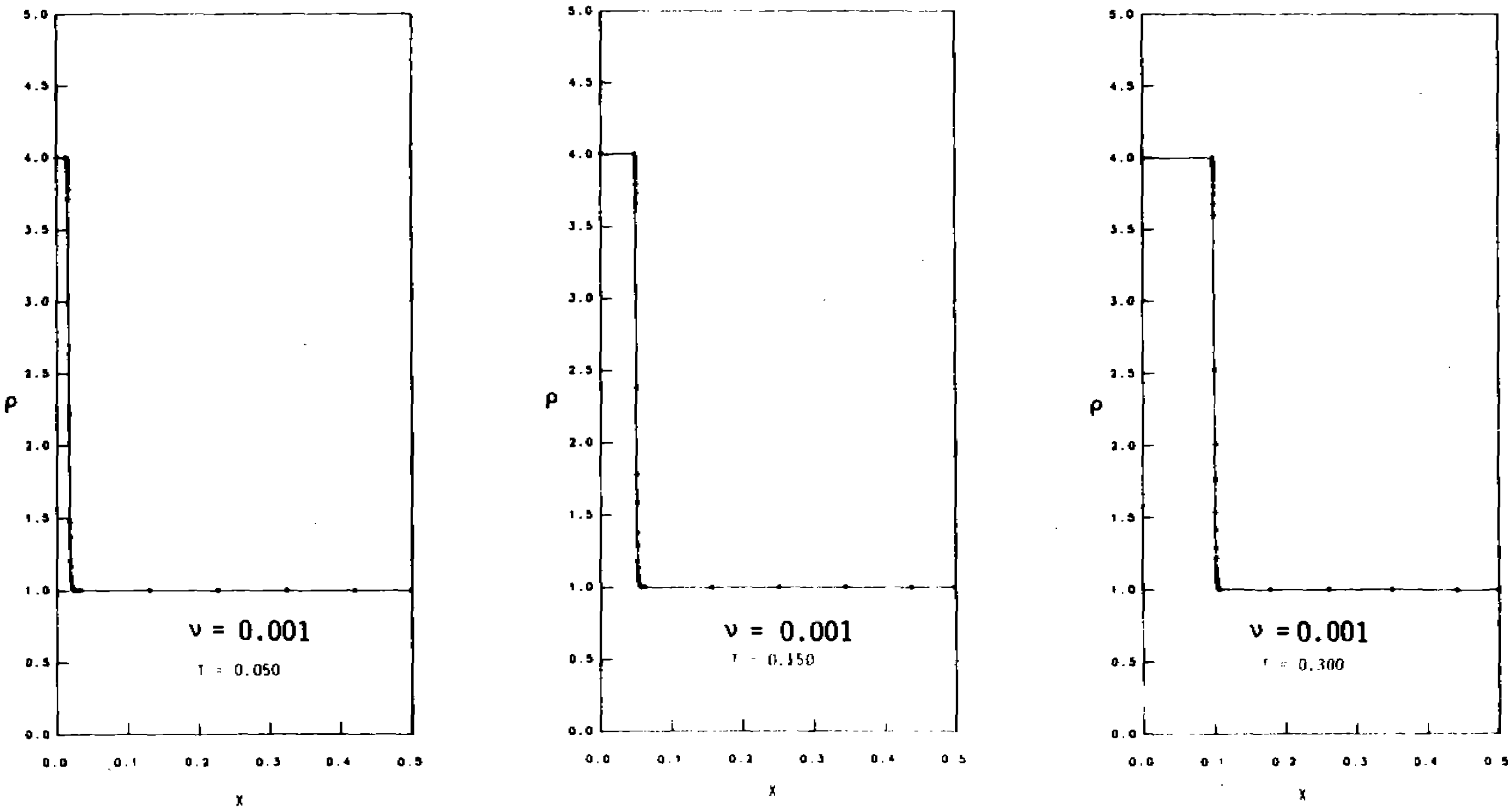


Figure 7. MFE solutions of the Navier-Stokes equations with slightly non-optimal node positions.

N-S

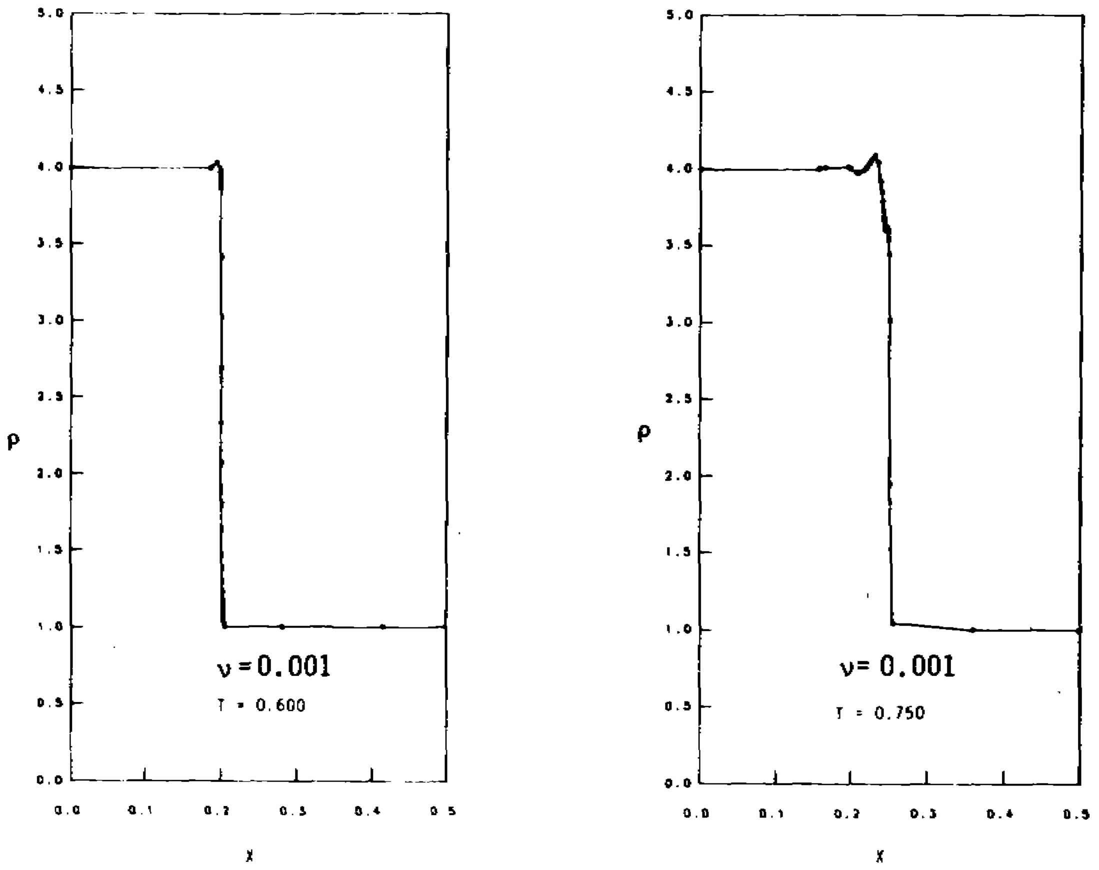


Figure 8. MFE solutions of the Navier-Stokes equations with slightly non-optimal node positions.

magnitudes on the order of $1/\nu$. Initial conditions which produce a doubly skewed wavefront profile are shown schematically in Figures 9 and 10. (The counterposed initial velocity fields are designed to create an evolving shock profile which is skewed in both the x- and y-components of velocity.) Boundary conditions are given by:

$$\begin{array}{rcll}
u(0,y) & = & u(1,y) = 0 & 0. \le y \le 1. \\
v_x(0, y) & = & v_x(1, y) = 0. & 0. \le y \le 1. \\
u(x, 1) & = & 0.2 \sin \pi x & 0. \le x \le 1. \\
u(x, 0) & = & -.2 \sin \pi x & 0. \le x \le 1. \\
v(x, 1) & = & -1. + 0.2 \cos \pi x & 0. \le x \le 1. \\
v(x, 0) & = & 1. + 0.2 \cos \pi x & 0. \le x \le 1 \quad .
\end{array}$$

The MFE nodes are fixed by zero-Dirichlet conditions along the top and bottom horizontal edges of the domain. The nodes are free to move vertically by symmetric boundary conditions along the left and right edges of the domain.

This problem can be solved readily by perhaps many PDE solution methods whenever ν assumes sufficiently large values. For example, a value of ν = 0.02 produces shock gradients on the order of 10^2.

The MFE method requires only an 8 x 8 grid to give reasonably accurate solutions to this problem, and Figures 11 and 12 show accurate MFE solutions on a 12 x 12 grid. Here Figure 11 presents an isometric view of the evolving profile of the y-component of velocity at t = 3.0, well after the shock has formed and after the wavefront has undergone significant shearing. The x-component of velocity is sufficiently sheared that a hidden line plot, which is not yet available, is required for easy interpretation by the naked eye. The MFE grid nodes have migrated extensively from their initial positions as can be seen in Figure 12 which represents the grid mesh projected onto the x-y plane at t = 3.0. Figures 13 and 14 present contour plots for selected constant values of u and v, respectively, at t = 3.0. It is evident from the magnitudes of shock gradients and from the regions of significant curvature which span nearly the entire domain that an alternative PDE method with a fixed grid may require on the order of 10^4, or more, grid nodes in order to achieve comparable degrees of accuracy in this problem.

This same basic problem can now be made to correspond to a much more demanding physical problem by letting ν = 0.002. Figure 15 shows an isometric view of the MFE solution on a 16 x 16 grid for this case. Shock gradients are now generated with magnitudes of several times 10^3. Before discussing these MFE results in detail, some general observations should be discussed: It is extremely unlikely that any existing PDE method using either a fixed grid or a less than optimal adaptive grid can accurately solve this test problem with fewer than 10^5-10^6 grid nodes. Grid aspect ratios frequently assume very large values (10^2-10^3). It should also be noted here that numerous inviscid solvers which are under development do not apply at all to this type of advection-diffusion problem because the Laplacian is an essential mathematical operator whose effects must be rigorously resolved in advection-diffusion PDE's. Because inviscid solvers do not generally solve PDE's which contain Laplacians, they generate shocks with gradient shapes and magnitudes that are governed exclusively by the grid spacing and/or by the purely numerical dissipative processes in the inviscid method, per se. Consequently, inviscid solvers have almost no chance of resolving those physical

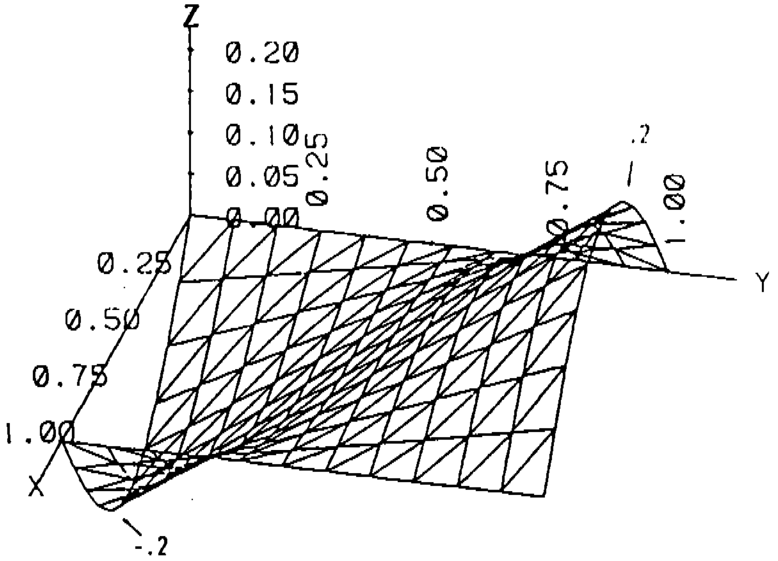


Figure 9. Plot of initial values of u in the 2-D Burger-like example on a 12 x 12 grid mesh.

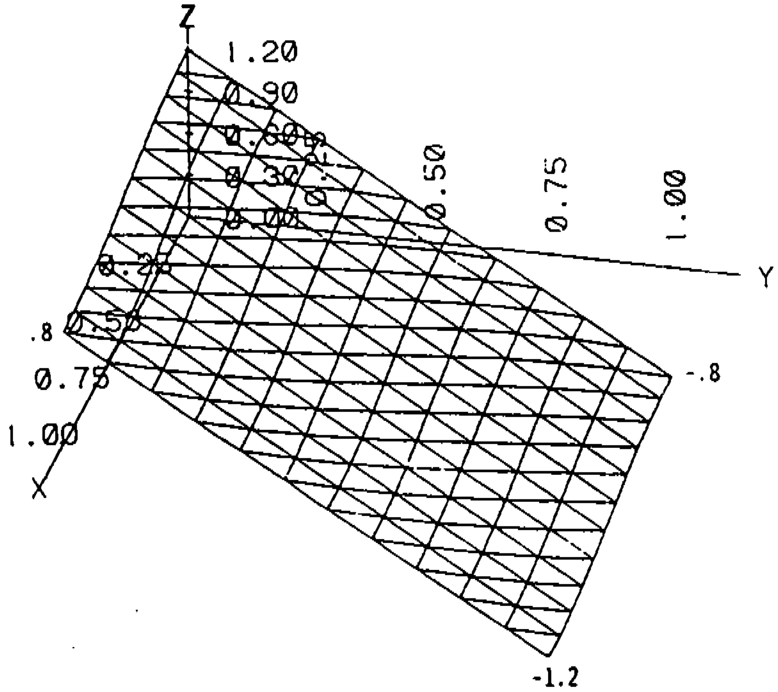


Figure 10. Plot of initial values of v in the 2-D Burger-like example on a 12 x 12 grid mesh.

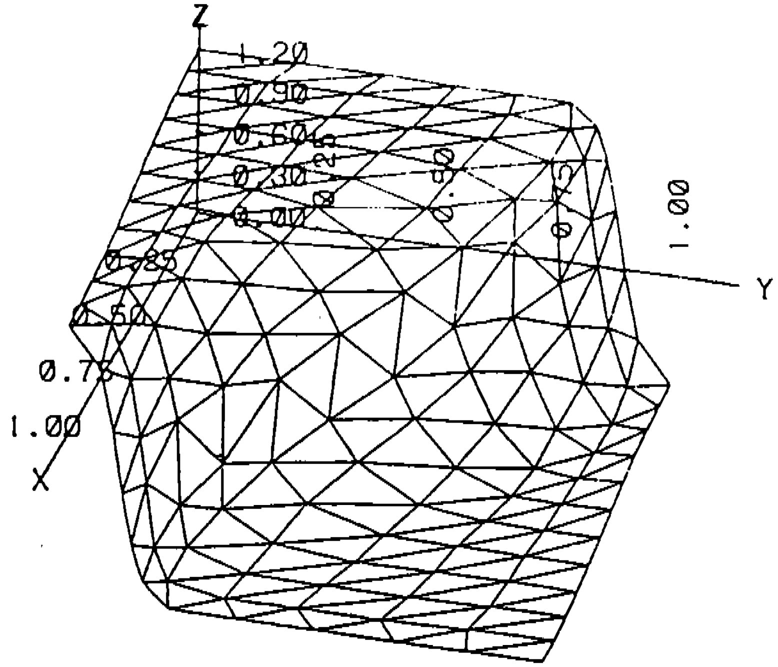


Figure 11. Isometric view of v at t = 3.0 in the 2-D Burger-like example on a 12 x 12 MFE grid.

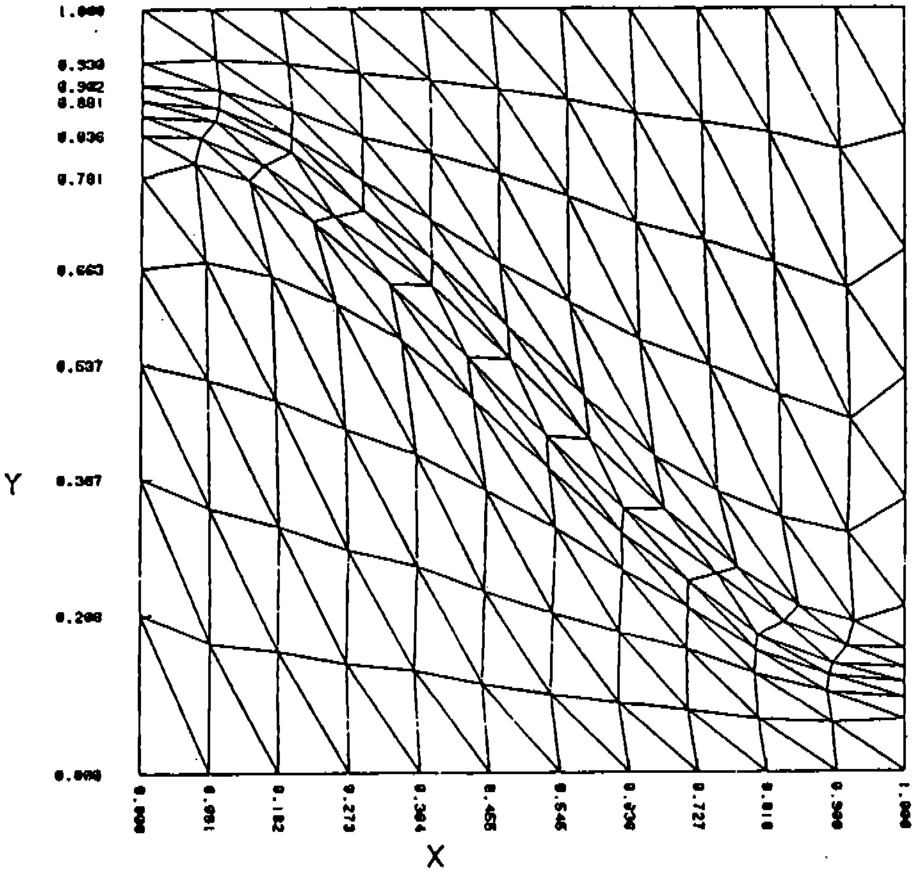


Figure 12. MFE grid projections on the x-y plane at t = 3.0 in the 2-D Burger-like example on a 12 x 12 grid.

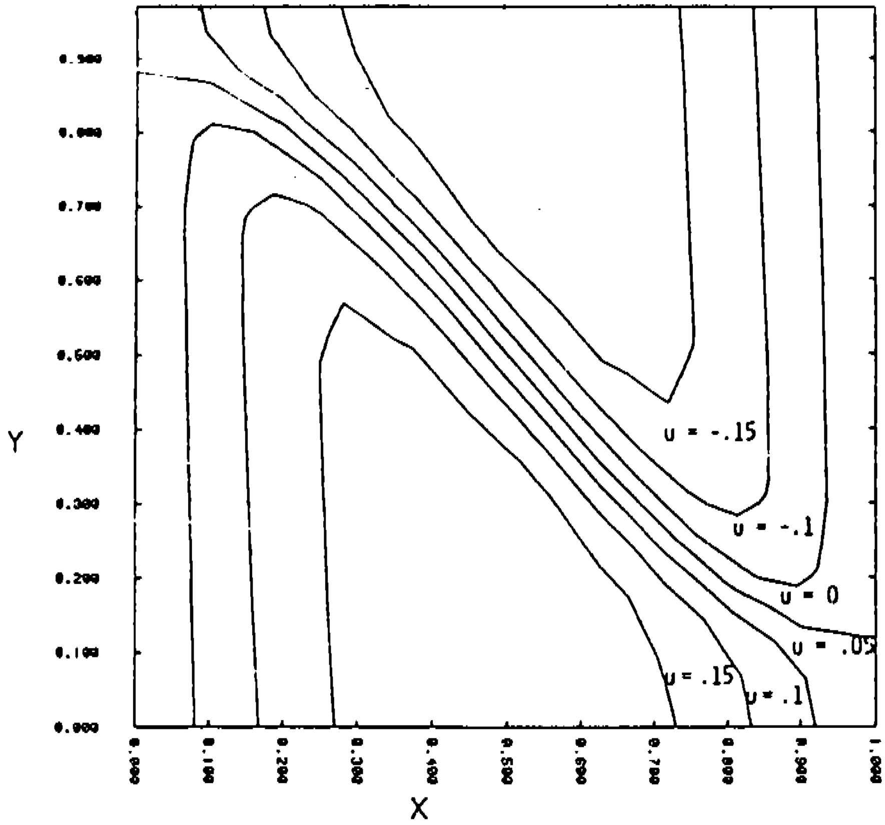


Figure 13. Contour plots of selected values of u at t = 3.0 in the 2-D Burger-like example on a 12 x 12 grid.

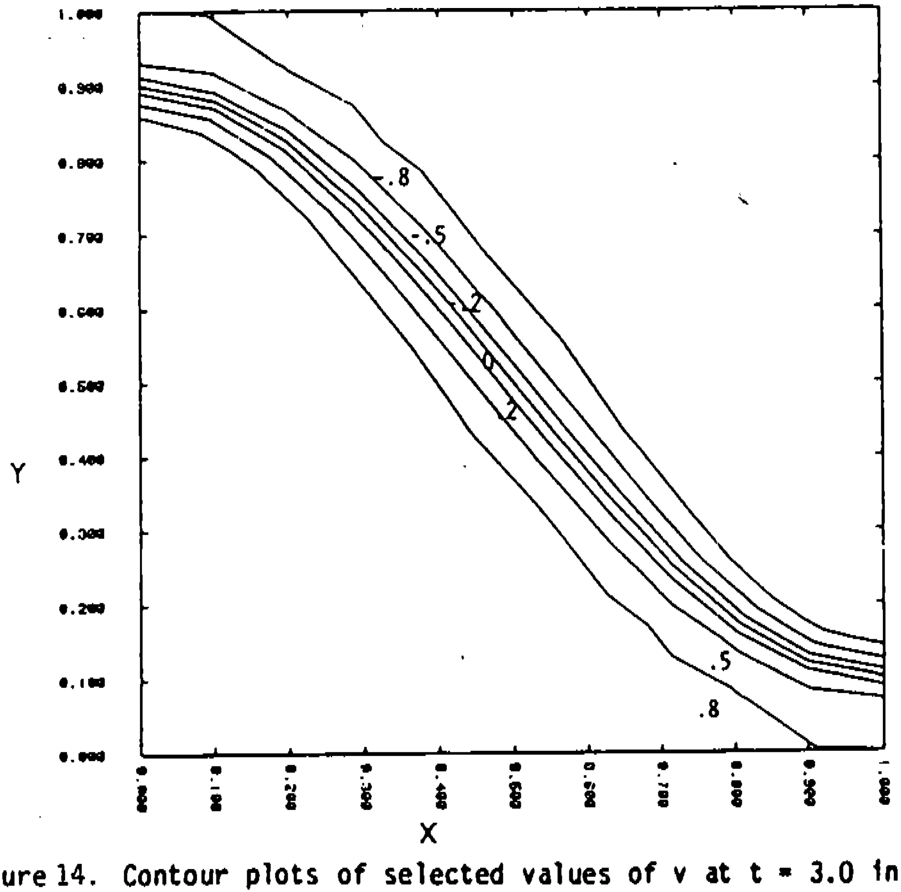


Figure 14. Contour plots of selected values of v at t = 3.0 in the 2-D Burger-like example on a 12 x 12 grid.

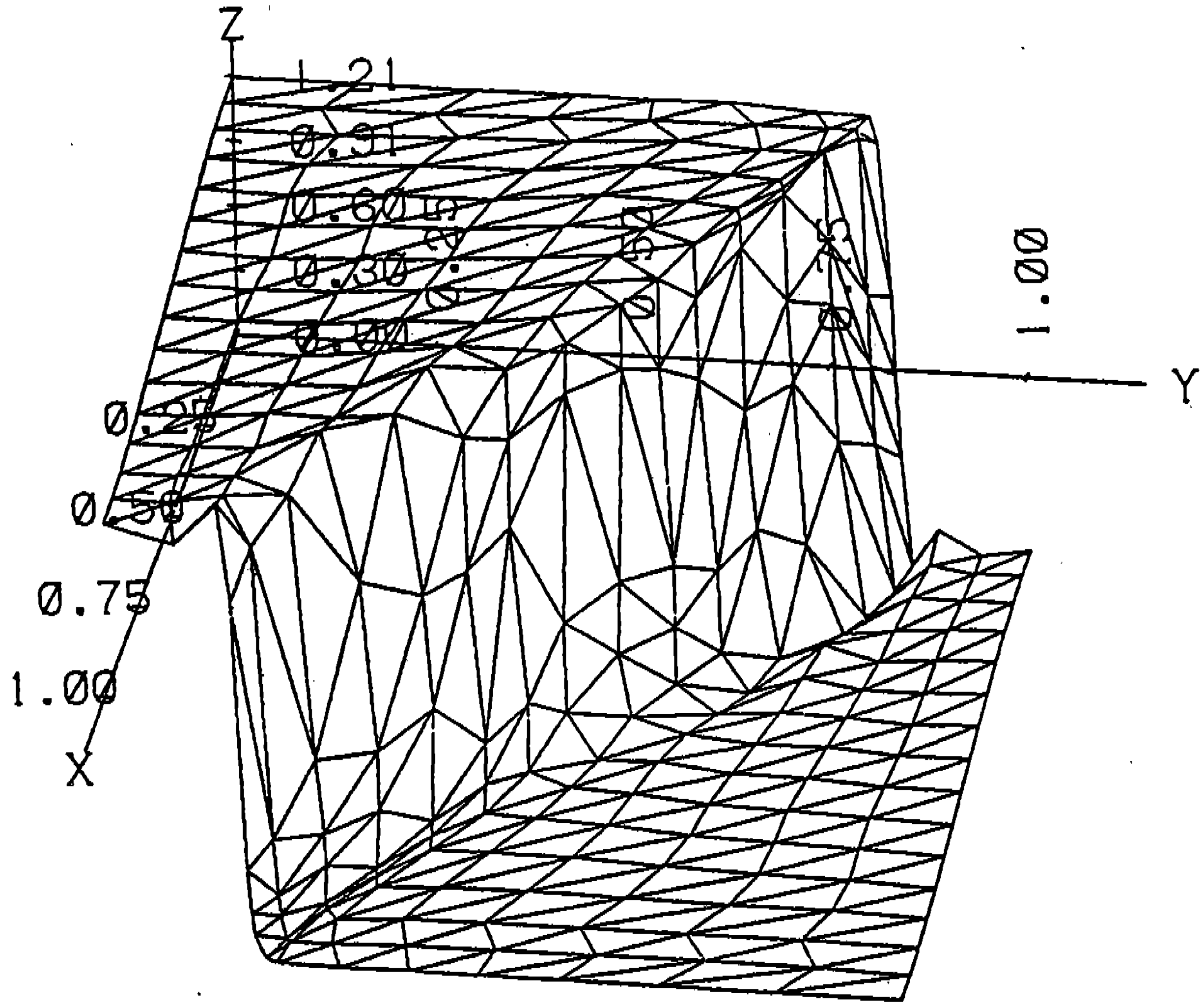


Figure 15. Isometric view of Burger's test example at t = 1.8 with ν = 0.002 on a 16 x 16 MFE mesh. Shock gradients have magnitudes of approximately 10^3 in this solution for u.

dissipation effects which are usually expressed by Laplacian operators and are present with fundamental physical significance in transport theory, hydrodynamics, plasma physics, continuum mechanics, and many other disciplines in the physical sciences. This critical discussion is not intended to denigrate the extensive research efforts on inviscid PDE solvers and/or fixed node PDE methods where they legitimately apply; but it does suggest that efforts to accommodate Laplacian operators in otherwise inviscid solution methods and efforts to investigate more optimal adaptive grid methods for use in many existing PDE methods which are applied to advectiondiffusion problems should now assume greatly increased significance. In the meantime, the MFE method is providing various clues to some of the significant new areas where mathematics research can profitably be intensified, as will be indicated below.

Early MFE results in 2-D are apparently continuing the trend which appeared in previous 1-D results. There, we saw that MFE solutions of both the Navier-Stokes and physically dissipative continuum mechanics equations in 1-D exhibited extremely high levels of simultaneous resolution of extremely disparate microscale and macroscale physical processes. While current 2-D MFE results exhibit similar promising features, numerous mathematical problems still require resolution in order to attain fully the desired levels of success in truly large-scale problems in 2-D. Clues to these problem areas can be seen in Figure 15.* For example, the irregularity of the grid triangles in the face of the shock could eventually prove to be troublesome. Similarly, the small oscillation at the base of the shock in this run is unsatisfactory, even though it can be eliminated in any number of ways. Extensive testing and analysis has indicated that the causal mechanisms underlying such mesh irregularities and oscillations in 2-D can be associated with: (i) time step and error control policies in the basic ODE integrator of Gear which is presently used, (ii) convergence properties of the linear solver, and (iii) limitations in the first-generation regularization functions. Each of these areas are worthy of continued intensive investigation.

4. Conclusions and Future Work.

- The MFE method can now be considered for large-scale applications in 2-D. Applications of the 2-D MFE method to airblast effects and structural impacts in continuum mechanics are in their early developmental phases.

- It is apparent that extensive new research is needed in numerous areas of applied mathematics in order for the MFE and other related adaptive grid PDE methods to reach their ultimate potential. Some of the key topical areas for additional research are:

  - ODE integrators for PDE applications.
  - Linear systems solvers for non-symmetric matrices which are not diagonally dominant. (Even the most advanced linear solvers for symmetric, diagonally dominant matrices converge too slowly for use in difficult advection-diffusion problems.) We have one new concept for such highly skewed matrix systems under development which is showing significant promise.
  - Boundary conditions for arbitrary domains.
  - Alternative MFE basis functions.
  - Alternative co-ordinate systems. (The MFE does not suffer from singularities at the origin in cylindrical and spherical co-ordinates.)
  - Alternative regularization schemes. (These powerful schemes are only now starting to be used in PDE solution methods.)

  The present research has made sufficient early advances in most of these areas to indicate the vast potential which lies ahead as new advances are made.

*It is apparent that these suggested mathematical problems will have to be resolved not only for the MFE method but also for most other advanced PDE methods which may seek to solve the difficult advection-diffusion equations which frequently arise in physical problems.

REFERENCES

1. Chapman, S. and T.G. Cowling, The Mathematical Theory of Non-Uniform Gases, Cambridge University Press, New York, 1960.

2. Hirschfelder, J.O., C.F. Curtiss and R.B. Bird, Molecular Theory of Gases and Liquids, John Wiley and Sons, New York, 1964.

3. Sod, G., "A Survey of Several Finite Difference Methods for Systems of Non-Linear Hyperbolic Conservation Laws," J. Comp. Phys., 27, p. 1-31, 1978.

4. Gelinas, RJ. and S.K. Doss, "The Moving Finite Element Method: A Semi-Automatic Solver for Diverse PDE Applications," Advances in Computer Methods for Partial Differential Equations--IV, pg. 230-239, edited by R. Vichnevetsky and R.S. Stepleman, Proceedings, Fourth IMACS International Symposium on Computer Methods for Partial Differential Equations, Lehigh University, Bethelehem, PA, June 30-July 2, 1981.

ELASTIC WAVE SCATTERING FROM CYLINDRICAL CAVITIES AND SOLID INCLUSIONS ANALYZED BY THE RESONANCE METHOD

P. P. Delsanto, J. D. Alemar, and E. Rosario
Department of Physics
University of Puerto Rico, Mayaguez, Puerto Rico 00709

A. Nagl, J. V. Subrahmanyam and H. Überall
Department of Physics
Catholic University of America, Washington, D.C. 20064

ABSTRACT. The formalism for a complete treatment of elastic wave scattering from cylindrical cavities and solid inclusions has been worked out for both perpendicular and oblique indice. Special cases, e.g., fluid cylinders enclosed in fluid or solid media, have already been numerically analyzed. Poles of scattering amplitudes in the complex frequency plane have been found, and were physically interpreted in terms of helical surface waves on the cylinders propagating in both the interior and the exterior medium. Dispersion, attenuation and refraction of these surface waves have been obtained.

I. INTRODUCTION. We consider the problem of scattering of obliquely incident elastic or acoustic waves from the surface of an infinite circular cylinder. There are four cases of interest:

(i) inside solid, outside solid
(ii) inside solid, outside fluid
(iii) inside fluid, outside solid
(iv) inside fluid, outside fluid

A complete analytical solution for case (i) was first obtained by White [1], whose work was later corrected by Lewis and Kraft [2]. The second case was discussed by Flax, Varadan and Varadan [3], and the appearance of resonance effects in the numerically calculated scattering cross section was noted. Case (iii) was treated by Solomon et al [4] for normal incidence. The present authors discuss in a recent paper [5] the refraction effects which take place upon the generation of surface waves in case (iv).

In the present paper we develop a general formalism that covers the four cases mentioned. This includes both formulas for the scattering amplitudes of the resonances and an analysis of their Watson poles in the complex frequency plane and their physical interpretation. We introduce a compact notation and a transformation gauge which simplify the solution considerably, making it suitable for computer coding.

As in White [1] we call 1 the inside region and 2 the outside region. We also use arabic subscripts (1,2) for longitudinal or compressional waves (p-waves) and roman subscripts (I,II) for transverse or shear waves (s-waves). Therefore, in general we call ϕ_2 the angle of the incident wave propagation vector (which we assume to be in the x-z plane, as in Fig. 1) with the x-axis; we call it, however, ϕ_{II} when we want to specify that the incident wave is an s-wave.

II. ELASTIC WAVE SCATTERING. The equation of motion for the particle displacement $\vec{u}$ in an elastic medium is

$$(\lambda + 2\mu)\vec{\nabla}\cdot\vec{\nabla}\,\vec{u} - \mu\vec{\nabla}\times(\vec{\nabla}\times\vec{u}) = \rho\partial^2\vec{u}/\partial t^2 \qquad (1)$$

where λ and μ are the Lamé constants of the medium and ρ is its density. The particle displacement $\vec{u}$ is obtained from a scalar (ψ) and a vector ($\vec{V}$) potential:

$$\vec{u} = \vec{\nabla}\psi + \vec{\nabla}\times\vec{V} \qquad (2)$$

where ψ contributes to p-waves and V to s-waves.

Inserting Eq. (2) into Eq. (1) leads to two wave equations for the potentials ψ and V:

$$\nabla^2\psi = (1/c_p^2)\frac{\partial^2\psi}{\partial t^2}$$

$$\nabla^2\vec{V} = (1/c_s^2)\frac{\partial^2\vec{V}}{\partial t^2}$$

where $c_p = \{(\lambda + 2\mu)/\rho\}^{1/2}$ is the speed of p-waves and $c_s = (\mu/\rho)^{1/2}$ is the speed of s-waves in the medium.

We define

$$p_i = k_i\sin\phi_i \qquad (3)$$

$$q_i = k_i\cos\phi_i \qquad (4)$$

where i=1,2,I,II and $k_i = \omega/c_i$ is the wavenumber. The angular frequency of the wave, which we assume to be monochromatic, is ω. Snell's law requires

$$\frac{\sin\phi_1}{c_1} = \frac{\sin\phi_2}{c_2} = \frac{\sin\phi_I}{c_I} = \frac{\sin\phi_{II}}{c_{II}} ,$$

leading to

$$p_1 = p_2 = p_I = p_{II} = p \tag{5}$$

A. P-WAVE INCIDENCE. The incident p-wave can be expressed as

$$\begin{aligned} \psi^i &= (1/\rho_2\omega^2)e^{i\vec{k}_2\cdot\vec{r}} = Ne^{ipz}e^{iq_2 r\cos\theta} \\ &= Ne^{ipz}\sum_{n=0}^{\infty} e_n i^n J_n(q_2 r)\cos n\theta \\ e_n &= 1\ (n=0),\ 2\ (n>0), \end{aligned} \tag{6}$$

where we have used the Jacobi-Anger formula [6].

We now introduce the following compact notation:

$$N = 1/\rho_2\omega^2 \tag{7}$$

$$\left.\begin{aligned} c_n &= Ne_n i^n e^{ipz}\cos(n\theta) \\ s_n &= Ne_n i^n e^{ipz}\sin(n\theta) \end{aligned}\right\} \tag{8}$$

$$c'_n = \partial c_n/\partial\theta \tag{9}$$

$J_n = J_n\ (qr)$ = Bessel function, $q=q_i$ and $i=1,2,I,II$ according to case

$H_n = H_n\ (qr)$ = Hankel function of first kind

We implicitly assume $\sum_{n=0}^{\infty}$ any time an index n occurs.

In this notation, Eq. (6) becomes

$$\psi^i = J_n c_n.$$

In the case of scattered p-waves an expansion of $J_n c_n$ can be used, but with coefficients R_n, T_n to be determined through a matching at the boundary (see Sec. D). Also in region 2, where the scattered (reflected) waves are outgoing, H_n must replace J_n, i.e.,

$$\begin{aligned} \psi^t &= T_n\ J_n\ c_n \qquad \text{for transmitted p-waves}, \\ \psi^r &= R_n\ H_n\ c_n \qquad \text{for reflected p-waves}. \end{aligned} \tag{10}$$

B. INCIDENT S-WAVES. The geometry in the case of incident s-waves as illustrated in Fig. 2, which shows that $\vec{V}$ has the components:

$$V_r = V(\cos\chi \sin\theta + \sin\chi \sin\phi \cos\theta)$$
$$V_\theta = V(\cos\chi \cos\theta - \sin\chi \sin\phi \sin\theta) \qquad (11)$$
$$V_z = -V \sin\chi \cos\phi$$

where

$$V = \frac{1}{\rho_2\omega^2}\exp(i\vec{k}_{II}\cdot\vec{r}) = J_n c_n \qquad (12)$$

since $\vec{\nabla}\times\vec{\nabla}\Phi = 0$ we can add to $\vec{V}$ and $\vec{\nabla}\Phi$. Now

$$\vec{\nabla}V = iq_{II}\cos\theta\, V\hat{r} - iq_{II}\sin\hat{\theta}\, V + ipV\hat{z} \qquad (13)$$

Therefore choosing

$$\Phi = -(\sin\chi\sin\phi V)/(iq_{II}) \qquad (14)$$

and adding $\vec{\nabla}\Phi$ to $\vec{V}$, it follows:

$$V_r = V\cos\chi \sin\theta$$
$$V_\theta = V\cos\chi \cos\theta \qquad (15)$$
$$V_z = -V\sin\chi/\cos\phi \;.$$

Equation 15 shows that any randomly polarized incident s-wave can be decomposed as a linear combination of a $\chi = 0$ wave (called SH or horizontally polarized) and a $\chi = \pi/2$ wave called SV or vertically polarized) with coefficients $\cos\chi$ and $\sin\chi$, respectively.

For SH-waves we have

$$V_r = V\sin\theta = -ni\, J_n s_n/q_{II}$$
$$V_\theta = V\cos\theta = -i\, J_n' c_n/q_{II} \qquad (16)$$
$$V_z = 0$$

and for SV-waves

$$V_r = V_\theta = 0$$
$$V_z = -V/\cos\phi = -J_n c_n/\cos\phi \qquad (17)$$

C. SCATTERED S-WAVES. The vector potential $\vec{V}$ can be expressed as

$$\vec{V} = \vec{\nabla}\Phi + \vec{\nabla}\Theta\times\hat{z} + \chi\hat{z} \qquad (18)$$

where Φ,Θ and χ are solutions of the Helmholts equation

$$(\nabla^2 + k_i^2)\Phi = 0, \qquad i = I,II \qquad (19)$$

In equation (18), $\vec{\nabla}(\partial\chi/\partial z)$ has been eliminated, since $\vec{\nabla}\times\vec{\nabla}f = 0$ for any function f. Also, instead of $k_i\chi$ we have simply written χ. Now

$$\begin{aligned} V_r &= \frac{\partial\Phi}{\partial r} + \frac{1}{r}\frac{\partial\Theta}{\partial\theta} \\ V_\theta &= \frac{1}{r}\frac{\partial\Phi}{\partial\theta} - \frac{\partial\Theta}{\partial r} \\ V_z &= \frac{\partial\Phi}{\partial z} + \chi. \end{aligned} \qquad (20)$$

Writing

$$\begin{aligned} \Phi &= (i/q_i)A_nJ_ns_n \\ \Theta &= (i/q_i)A_nJ_ns_n \\ \chi &= c_nJ_ns_n \end{aligned} \qquad (21)$$

and using recurrence relations for Bessel functions we get for the components of $\vec{V}$:

$$\begin{aligned} V_r &= -A_niJ_{n+1}s_n \\ V_\theta &= A_niJ_{n+1}c_n \\ V_z &= B_nJ_ns_n\ , \end{aligned} \qquad (22)$$

where A_n and B_n are coefficients to be determined through the matching of the inside and outside region at the cylinder surface.

In the case of SV-waves we put

$$\begin{aligned} \Phi &= (i/q_i)A_nJ_nc_n \\ \Theta &= (-i/q_i)A_nJ_ns_n \end{aligned}$$

and arrive at Eq. 22 with s_n and c_n interchanged. To treat both cases simultaneously we use Eq. 22 with the understanding that for SV-waves s_n and c_n must be interchanged. Also, whenever a difference of sign occurs we write $\pm$ or $\mp$ with the lower sign applying to the case of incident SV-waves.

D. THE MATCHING. At the surface of the cylinder r=a, we impose the following six boundary conditions:

$$\begin{aligned} u_j^t - u_j^r &= u_j^i \\ \mu_1\sigma_{rj}^t - \mu_2\sigma_{rj}^r &= \mu_2\sigma_{rj}^i \end{aligned} \qquad (23)$$

with j = r, θ, z; i.e., we require the continuity for the three components of the displacement $\vec{u}$ and stress tensor σ_{rj} of the incident, transmitted and reflected waves. This gives a system of six equations in the six unknowns R_n, T_n, A_n, B_n, F_n, G_n for the three cases of

(i) incident p-wave
(ii) incident SH-wave
(iii) incident SV-wave

The system, in the more general case (solid inside and outside) is given in Table 1. In the first six columns are the coefficients of the six unknowns. In the last three columns are the inhomogeneous terms in the three cases. If inside (or outside) we have a fluid, then $\mu = 0$, Eq. 2 and 3 expressing continuity of u_θ and u_z do not apply, and $A_n = B_n = 0$ ($F_n = G_n = 0$ for fluid outside).

This system of equations may be solved in a straightforward manner, e.g. by using Cramer's rule. In the following, we shall present numerical results for its solutions in some particular cases.

III. RIGID INCLUSION IN ELASTIC MEDIUM. We first specialize the above solution to the case of normal incidence of p or s waves on a rigid cylindrical inclusion. The determinant of the system of Table 1 which, after the case of Cramer's rule, also becomes the common denominator of the coefficients R_n through G_n, reduces in this special case to

$$\det = \begin{vmatrix} D_{11} & D_{12} \\ D_{31} & D_{32} \end{vmatrix} \qquad (24)$$

where a is the cylinder radius, and

$$\begin{aligned}
D_{11} &= k_2a\ H_{n-1}(k_2a)-nH_n\ (k_2a) \\
D_{12} &= n\ H_n\ (k_{II}a) \\
D_{31} &= n(n+1)\ H_n(k_2a)-nk_2a\ H_{n-1}(k_2a) \\
D_{32} &= k_{II}a\ H_{n-1}(k_{II}a)+\left[\tfrac{1}{2}(k_{II}a)^2-n(n+1)\right]\ H_n(k_{II}a).
\end{aligned} \tag{25}$$

The poles of the scattering amplitude are given by the zeros of Eq. (24) in the complex frequency plane. These were obtained by us numerically, and are shown in Fig. 3 for the case of the elastic medium being aluminum, plotted in the complex plane of the variable k_2a . These zeros are seen to have negative imaginary parts. Zeros with given value of n (as indicated) are joint by dashed lines. However, as was shown in connection with sound scattering from a sphere [7], physical interpretations are obtained if one considers the zeros to be grouped in "layers", connected by solid lines in Fig. 3 and labeled by the integer ℓ as indicated. It was shown in [7] that the residues of the scattering amplitude, summed over the poles along each solid line in Fig. 3, synthesize a surface wave which for the case of a cylindrical inclusion, propagates over the cylinder surface on a circular path for normal incidence, or as a helical path for oblique incidence [5]. The dispersion curves of these waves are now being obtained using standard methods [8]. In addition to these "bulk type" (p or s) surface waves, a "Rayleigh type" surface wave is also found from our pole calculations, similar to the case of a spherical cavity as discussed by Norwood and Miklowitz [9]. In addition, corresponding poles and surface waves were obtained for an empty cavity in aluminum.

IV. FLUID-FILLED CAVITY IN ELASTIC MEDIUM. For the case of normal incidence of p or s waves on a fluid-filled cylindrical cavity, the corresponding determinant from Table 1 is

$$\det = \begin{vmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & 0 \end{vmatrix} \tag{26}$$

where, in addition to Eq. 25,

$$\begin{aligned}
D_{13} &= n\ J_n(k_1a)-k_1a\ J_{n-1}(k_1a), \\
D_{21} &= \left[n(n+1)-\tfrac{1}{2}(k_{II}a)^2\right] H_n(k_2a)-k_2a\ H_{n-1}(k_2a), \\
D_{22} &= n\ k_{II}a\ H_{n-1}(k_{II}a)-n(n+1)H_n(k_{II}a) \\
D_{23} &= \tfrac{1}{2}(\rho_1/\rho_2)\ (k_{II}a)^2\ J_n\ (k_{II}a).
\end{aligned} \tag{27}$$

The poles, obtained from the zeros of Eq. (26) for a water-filled cavity in Aℓ , were found to be numerically close to those for p,s and Rayleigh-type surface waves of the above-mentioned empty cavity.

In addition, pole layers were found that lay very close to the real axis of the complex k_2a plane; these correspond to internal surface waves that propagate in the filler fluid. They appear in the form of resonances in the scattering amplitude when the latter is plotted vs. frequency, as seen in Fig. 4. This figure shows, for backscattering, the modulus of the p→p scattering amplitude R_n for n = 0 through 5. The resonances interfere destructively with the background amplitude $R_n^{(o)}$ for an empty cavity, and if the latter background is subtracted, the resulting pure resonances of the internal surface waves appear in Fig. 5.

The resonance frequencies, which on the k_2a scale we designate by $x_{n\ell}$, can be read off this figure, and provide us with the real parts of the complex-frequency poles. The imaginary parts are negligible to first order, indicating a minimal attenuation of these internal surface waves. Note that the same resonance frequencies are obtained for p→p and s→s scattering as well as for p→s and s→p scattering ("mode conversion") since physically, they all originate from the eigenfrequencies of the internal fluid.

The phase and group velocities of the internal fluid can now be obtained from $x_{n\ell}$ as follows [4,5]:

$$c_{ph} = (c_2/n)x_{n\ell} \ , \tag{28}$$

$$c_{gp} = c_2 \left[\partial(k_2a)/\partial n \right]_{k_2a = x_{n\ell}} \ . \tag{29}$$

In Fig. 6, we show the corresponding phase velocity dispersion curves of the internal sufrace waves as obtained from Eq. 8.

These dispersion curves help in determining the refraction angles that appear during the generation of a surface wave by an externally incident plane wave, as outlined in Reference [5].

ACKNOWLEDGEMENTS. Research support at the University of Puerto Rico, Mayaguez, by the U.S. Army Research Office under Grant No. DAAG 29-81-G0016 is gratefully acknowledged. We wish to thank Dr. S. G. Solomon for some helpful contributions.

REFERENCES.

[1] R. M. White, "Elastic wave scattering at a cylindrical discontinuity in a solid", J. Acoust. Soc. Am. 30, 771 (1958).

[2] S. T. Lewis and D. W. Kraft, "Mode conversion relation for an elastic wave scattered by a cylindrical obstacle in a solid", J. Acoust. Soc. Am. 56, 1899 (1974).

Table 1. Coefficients of $R_n \ldots G_n$; coefficients of incident p, SH, SV waves.

	R_n	T_n	A_n	B_n	F_n	G_n	P	SH	SV
u_r	H'_n	$\bar{J}'_n$	$p\bar{J}_{\overline{n+1}}$	$\pm\frac{n}{r}\bar{J}_{\bar{n}}$	$pH_{\overline{n+1}}$	$\pm\frac{n}{r}H_{\bar{n}}$	J'_n	$-tJ'_{\bar{n}}$	$\frac{nJ_{\bar{n}}}{r\cos\phi}$
u_θ	$\mp\frac{n}{r}H_n$	$\mp\frac{n}{r}\bar{J}_n$	$p\bar{J}_{\overline{n+1}}$	$-\bar{J}'_{\bar{n}}$	$pH_{\overline{n+1}}$	$-H'_{\bar{n}}$	$-\frac{n}{r}J_n$	$\frac{n}{r}tJ_{\bar{n}}$	$\frac{J'_{\bar{n}}}{\cos\phi}$
u_z	pH_n	$p\bar{J}_n$	$\bar{J}'_{\overline{n+1}}+\frac{1\pm n}{r}\bar{J}_{\overline{n+1}}$	0	$H'_{\overline{n+1}}+\frac{1\pm n}{r}H_{\overline{n+1}}$	0	pJ_n	$q_{II}J_{\bar{n}}$	0
σ_{rr}	$2\mu_2\{(\frac{n^2}{r^2}-q_2^2)H_n - \frac{1}{r}H'_n\}-\lambda_2K_2^2H_n$	$2\mu_1\{(\frac{n^2}{r^2}-q_1^2)\bar{J}_n - \frac{1}{r}\bar{J}'_n\}-\lambda_1K_1^2\bar{J}_n$	$2p\bar{J}'_{\overline{n+1}}\mu_1$	$\pm\frac{2n}{r}\bar{J}^-_{\bar{n}}\mu_1$	$2pH'_{\overline{n+1}}\mu_2$	$\pm\frac{2n}{r}H^-_{\bar{n}}\mu_2$	$2\mu_2\{(\frac{n^2}{r^2}-q_2^2)J_n - \frac{1}{r}J'_n\}-\lambda_2K_2^2J_n$	$2t\mu_2\{\frac{1}{r}J'_{\bar{n}} + (q_{II}^2-\frac{n^2}{r^2})J_{\bar{n}}\}$	$\frac{2nJ^-_{\bar{n}}}{r\cos\phi}\mu_2$
$\sigma_{r\theta}$	$\mp\frac{2n}{r}H^-_n\mu_2$	$\mp\frac{2n}{r}\bar{J}^-_n\mu_1$	$p\mu_1(\bar{J}'_{\overline{n+1}} - \frac{1\pm n}{r}\bar{J}_{\overline{n+1}})$	$\{(q_I^2-\frac{2n^2}{r^2})\bar{J}_{\bar{n}} + \frac{2}{r}\bar{J}'_{\bar{n}}\}\mu_1$	$p\mu_1(H'_{\overline{n+1}} - \frac{1\pm n}{r}H_{\overline{n+1}})$	$\{(q_{II}^2-\frac{2n^2}{r^2})H_{\bar{n}} + \frac{2}{r}H'_{\bar{n}}\}\mu_2$	$-\frac{2n}{r}J^-_n\mu_2$	$\frac{2n}{r}tJ^-_{\bar{n}}\mu_2$	$\{(\frac{2n^2}{r^2}-q_{II}^2)J_{\bar{n}} - \frac{2}{r}J'_{\bar{n}}\}\frac{\mu_2}{\cos\phi}$
σ_{rz}	$2pH'_n\mu_2$	$2p\bar{J}'_n\mu_1$	$\{\pm\frac{n}{r}\bar{J}'_{\overline{n+1}}+(p^2-q_I^2 + \frac{n(n+2\mp1)}{r^2})\bar{J}_{\overline{n+1}}\}\mu_1$	$\pm\frac{pn}{r}\bar{J}_{\bar{n}}\mu_1$	$\{\pm\frac{n}{r}H'_{\overline{n+1}}+(p^2-q_{II}^2 + \frac{n(n+2\mp1)}{r^2})H_{\overline{n+1}}\}\mu_2$	$\pm\frac{pn}{r}H_{\bar{n}}\mu_2$	$2pJ'_n\mu_2$	$-\frac{\mu_2}{q_{II}}\{\frac{2n^2}{r^3}J_{\bar{n}} + (p^2-q_{II}^2-\frac{1}{r^2})J'_{\bar{n}}\}$	$\frac{n}{r}K_{II}tJ_{\bar{n}}\mu_2$

[3] L. Flax, V. K. Varadan, and V. V. Varadan, "Scattering of an obliquely incident acoustic wave by an infinite cylinder," J. Acoust. Soc. Am. 68, 1832 (1980).

[4] A. J. Haug, S. G. Solomon, and H. Überall, "Resonance theory of elastic wave scattering from a cylindrical cavity", J. Sound. Vib. 57, 51 (1978); S. G. Solomon, Ph D Thesis, Catholic University of America (1979); S. G. Solomon, H. Überall, and K. B. Yoo, "Mode conversion and resonance scattering of elastic waves from a cylindrical fluid-filled cavity", Acustica (submitted).

[5] A. Nagl, H. Überall, P. P. Delsanto, J. D. Alemar, and E. Rosario, "Refraction effects in the generation of helical surface waves on a cylindrical obstacle", Wave Motion (in press).

[6] See, e.g., W. Magnus, and F. Oberhettinger, Formulas and theorems for the functions of mathematical physics, Chelsea Publ. Co., New York (1949) p. 18.

[7] H. Überall, G. C. Gaunaurd, and J. D. Murphy, "Acoustic surface wave pulses and the ringing of resonances", J. Acoust. Soc. Am. 72, 1014 (1982).

[8] See, e.g., E. Hönl, A. W. Maue, and K. Westpfahl, "Theory of diffraction"; in Handbook of Physics XXV/1 (S. Flügge, ed.); Springer , Berlin (1961).

[9] F. R. Norwood and J. Miklowitz, "Diffraction of transient elastic waves by a spherical cavity", J. Appl. Mech. 34, 735 (1967).

FIGURE CAPTIONS

Fig. 1. Geometry of p-wave incidence on a cylindrical obstacle in an elastic medium.

Fig. 2. Geometry of s-wave incidence on a cylindrical obstacle in an elastic medium.

Fig. 3. Poles of the scattering amplitude for a rigid cylindrical inclusion in aluminum, plotted in the complex frequency plane.

Fig. 4. Modulus of $p \rightarrow p$ backscattering amplitude vs. frequency for normal incidence on a water-filled cavity in aluminum.

Fig. 5. As in Fig. 4, after subtraction of an empty-cavity background amplitude.

Fig. 6. Dispersion curves of internal surface waves in a water-filled cylindrical cavity in aluminum.

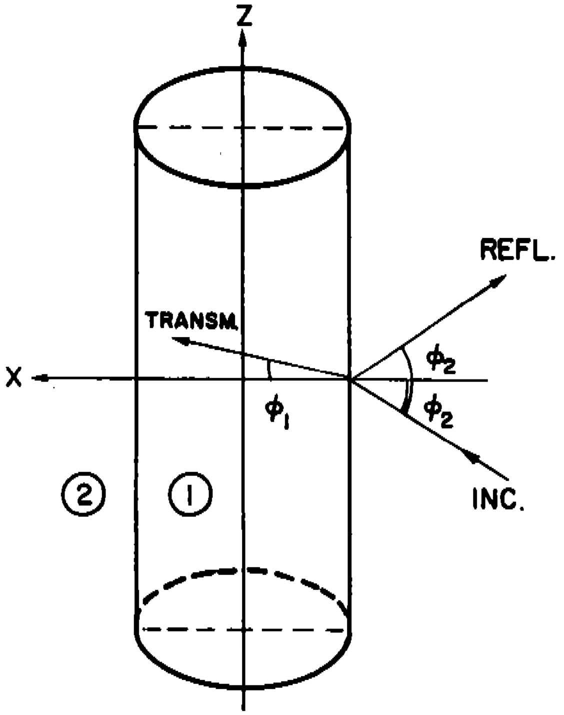


Fig. 1

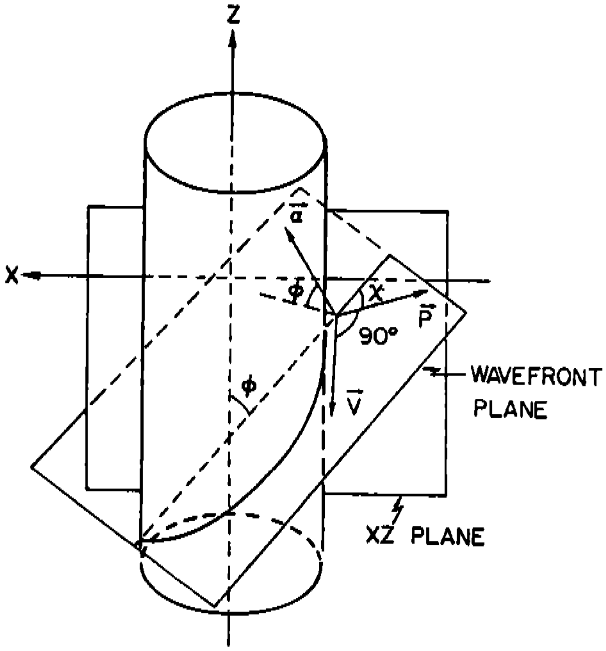


Fig. 2

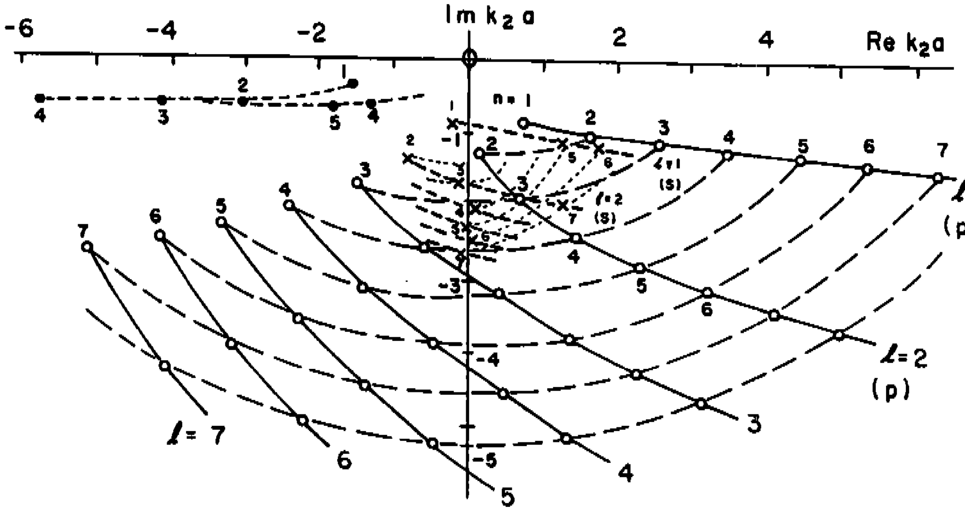


Fig. 3

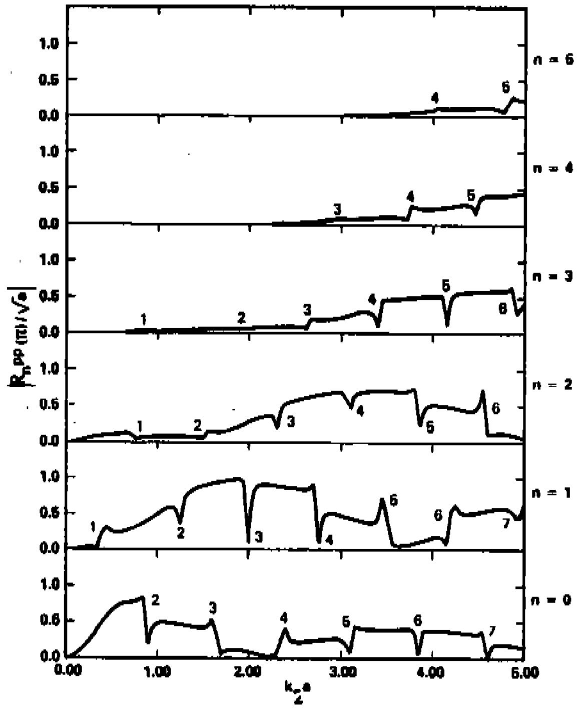


Fig. 4

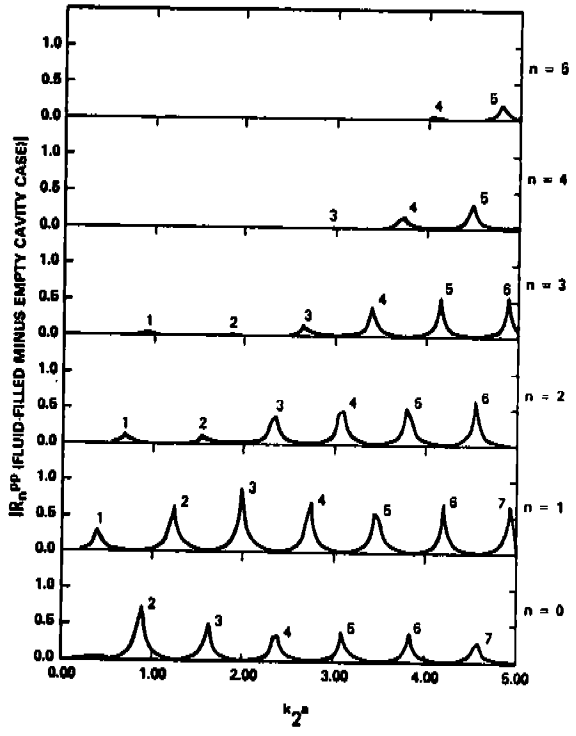
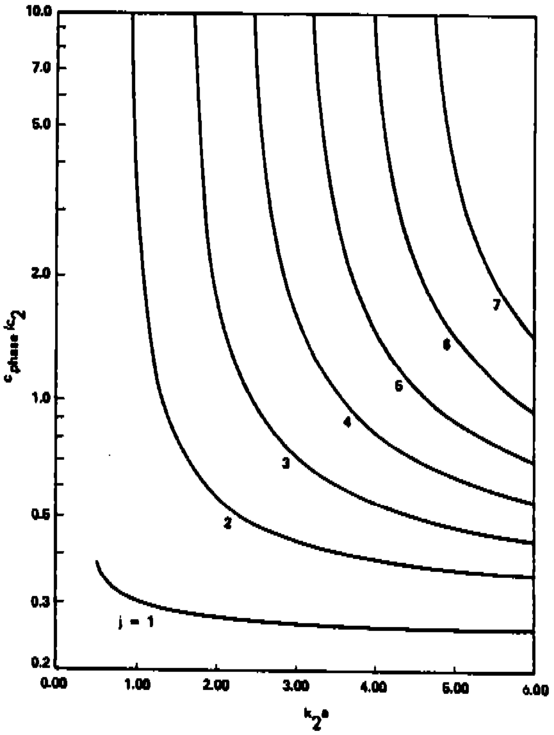


Fig. 5

Fig. 6



STABILITY OF PLANE NEAR-EQUIDIFFUSIONAL FLAMES WITHOUT INVOKING THE CONSTANT-DENSITY APPROXIMATION

T. Jackson and A. Kapila
Rensselaer Polytechnic Institute
Troy, New York 12181

INTRODUCTION. Near-equidiffusional flames (NEFs) are characterized by near-unity Lewis numbers, near-adiabatic flame temperatures and near-uniform enthalpies. These flames are special three-dimensional solutions of the combustion equations and provide a convenient framework for the theoretical study of a number of flame phenomena. An extensive discussion of NEFs appears in Chapters 2 and 8 of Buckmaster and Ludford (1982), where a set of reduced equations governing these flames has also been derived. This paper is concerned with a study of the linearized stability problem associated with the plane NEFs.

GOVERNING EQUATIONS. From the full combustion equations involving Arrhenius kinetics, a set of reduced equations appropriate to the NEFs can be derived in the limit of large activation energy $(\varepsilon \to 0)$, under the characterizing assumptions

$$L^{-1} = 1 + \varepsilon\ell/\alpha \quad , \quad H = T + \alpha Y = 1 + \alpha + \varepsilon h \quad ,$$

where α is the heat of reaction, L the Lewis number, T the temperature, Y the mass fraction of the reactant and H the enthalpy. All quantities are suitably nondimensionalized. Additional quantities appearing below are the density ρ, the velocity $\underline{v}$ and the pressure p (measuring deviations from the ambient pressure of unity on the $O(M^2)$ scale, M being a representative Mach number of the flame). Lengths are measured on the scale of the preheat zone and a frame of reference at rest in the laboratory is chosen. Then, to leading order in ε, the governing equations become

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\underline{v} = 0 \quad , \quad \rho\frac{D\underline{v}}{Dt} + \nabla p = \frac{1}{3}\nabla(\nabla\cdot\underline{v}) + \nabla^2\underline{v} \ , \tag{1}$$

$$\rho\frac{DT}{Dt} - \nabla^2 T = 0 \quad \text{in the unburnt region,} \tag{2}$$
$$T = 1 + \alpha \quad \text{in the burnt region,}$$

$$\rho\frac{Dh}{Dt} - \nabla^2 h = (\ell/\alpha)\nabla^2 T \quad , \quad \rho T = 1 \ , \tag{3a,b}$$

where

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \underline{v}\cdot\nabla$$

and for simplicity, Prandtl number is set at unity. These equations are supplemented by appropriate boundary conditions and jump conditions across the interface separating the burnt and unburnt regions.

The solution of the NEF problem in full generality is difficult, since nonlinear partial differential equations on either side of a possibly moving interface are involved. Most of the progress has been made on the basis of the constant-density approximation (CDA) which ignores density changes suffered by the gas in passing through the flame, by taking $\rho = 1$ and omitting (3b). The basic system then reduces to the linear equations (2) and (3a), in which $\underline{v}$ is taken to be a prescribed solution of equations (1). Although the approximation is drastic, it does reproduce qualitatively several features of real flames.

THE STABILITY PROBLEM. Under the CDA, the linearized stability problem associated with a plane flame leads to a system of ordinary differential equations with constant coefficients, which can be solved readily to yield the relationship between the growth rate ω and the transverse wave number k. This relationship leads to the dotted-line portion of the stability diagram of Fig. 1. (Details can be seen in Buckmaster and Ludford (1982), but the original work is due to Sivashinsky.) The neutral stability curves in the ℓk-plane produce a stable region in the middle, flanked on either side by regions of instability. Bifurcation to cellular structure occurs as the left stability boundary is crossed, while passage across the right stability boundary yields a pulsatile flame.

The CDA problem can be derived rigorously from the full NEF problem as the leading approximation in the limit $\alpha \to 0$. The purpose of this paper is to analyze the stability of a plane flame without involving the $\alpha \to 0$ limit. After considerable manipulation, the stability problem can be reduced to the equations

$$T'' - T' - \left(k^2 + \frac{\omega}{1 + \alpha e^x}\right)T - (m - k^2)\alpha e^x = 0 ,$$

$$h'' - h' - \left(k^2 + \frac{\omega}{1 + \alpha e^x}\right)h + (m - k^2)\ell(1 + x)e^x +$$

$$+ \frac{\ell}{\alpha}\left(T' + \frac{\omega}{1 + \alpha e^x} T + \alpha m e^x\right) = 0 ,$$

$$m''' - \frac{1 - \alpha e^x}{1 + \alpha e^x} m'' - \left(k^2 + \frac{\omega}{1 + \alpha e^x}\right)m' -$$

$$- \frac{\alpha e^x}{1 + \alpha e^x}\left(\frac{\omega}{1 + \alpha e^x} - \frac{1}{3} k^2\right)m +$$

$$+ \frac{1}{1 + \alpha e^x}\left(\frac{2\omega\alpha e^x}{(1 + \alpha e^x)^2} + \frac{1}{3} k^2\right)T' +$$

$$+ \frac{\omega}{(1 + \alpha e^{x})^2} (\frac{1}{3} k^2 - \frac{2\alpha^2 e^{2x}}{(1 + \alpha e^{x})^2}) T - \frac{k^2}{1 + \alpha e^{x}} p + \frac{\omega k^2 \alpha e^{x}}{(1 + \alpha e^{x})^2} = 0 ,$$

$$p' - \frac{1}{3}(k^2 + \frac{\omega}{(1 + \alpha e^{x})^2}) T - \frac{1}{3}(1 + \frac{\omega}{1 + \alpha e^{x}}) T' -$$

$$- (1 + \alpha e^{x}) m'' + (1 - \frac{4}{3} \alpha e^{x}) m' + (\omega - \frac{2}{3} \alpha e^{x} + k^2 (1 + \alpha e^{x})) m +$$

$$+ (\frac{\omega^2}{1 + \alpha e^{x}} + k^2 (\omega + \frac{1}{3} \alpha e^{x})) = 0 ,$$

subject to the boundary conditions

$$T = h = p = 0 \quad , \quad m = - \omega \quad \text{at} \quad x = - \infty \quad ,$$

and, at $x = 0$,

$$T = 0 \quad , \quad T' - \frac{\alpha}{2} h = 0 \quad , \quad h' + \frac{\ell}{\alpha} T' - \lambda h = 0 \quad ,$$

$$m + \frac{\omega}{1 + \alpha} - \frac{m'}{\lambda} + \frac{k(\lambda + k)}{\lambda\{k(1 + \alpha) - \omega\}} \{p - \frac{4}{3} \alpha (m + \frac{h}{2})\} = 0 \quad ,$$

$$m + \frac{\omega}{1 + \alpha} + \frac{k(\lambda^2 - k^2)}{\lambda^2 \{k(1 + \alpha) - \omega\}} \{p - \frac{4}{3} \alpha (m + \frac{h}{2})\} - \frac{m''}{\lambda^2} -$$

$$- \frac{\alpha m'}{\lambda^2 (1+\alpha)} + \frac{\omega T'}{\lambda^2 (1 + \alpha)^2} - \frac{\alpha \omega T}{\lambda^2 (1 + \alpha)^3} - \frac{k^2 \alpha}{\lambda^2 (1 + \alpha)} = 0 \quad .$$

Here,

$$\lambda = \frac{1}{2} - \frac{1}{2} \{1 + 4 k^2 + \frac{4\omega}{1 + \alpha}\}^{1/2} \quad ,$$

and primes denote differentiation with resppect to x . The reduction of an infinite domain to a semi-infinite domain $(x < 0)$ is achieved by being able to solve the problem analytically in the burnt region $(x > 0)$. The variable T, h, m and p denote, respectively, the purturbations in temperature, enthalpy, mass flux and pressure. Solutions of the form

$$\phi (x, y, t) = \phi (x) e^{iky+\omega t}$$

have been sought, thereby limiting the perturbations to only two spatial dimensions, for simplicity. Three-dimensional extension can be made at minor additional expense.

Unlike the CDA problem, the non-CDA equations have to be treated numerically, and this was done by appending a root finder based on Muller's procedure to the boundary-value solver COLSYS. Numerical results for $\alpha = 1$ are shown in Fig. 1 by the continuous curves through the actually computed points (denoted by asterisks). The right stability boundary does not appear to differ much from the corresponding CDA curve. The left stability boundary, on the other hand, deviates substantially from its CDA counterpart for small k, veering sharply to the right to narrow the stability region for long waves. Unfortunately, the non-CDA results are not yet available for the approximate range $5 < \ell < 10.5$, as the numerics breaks down in that region. The corresponding problem is currently under investigation. For large k, an asymptotic analysis shows that the non-CDA results are the same as the CDA results, to leading order. This fact was confirmed by the numerics.

To summarize, the effect of accounting for variation in density appears to be most pronounced for long waves, whose region of stability is substantially reduced.

REFERENCES

J. D. Buckmaster, and G. S. S. Ludford, **Theory of Laminar Flames**, Cambridge University Press, 1982.

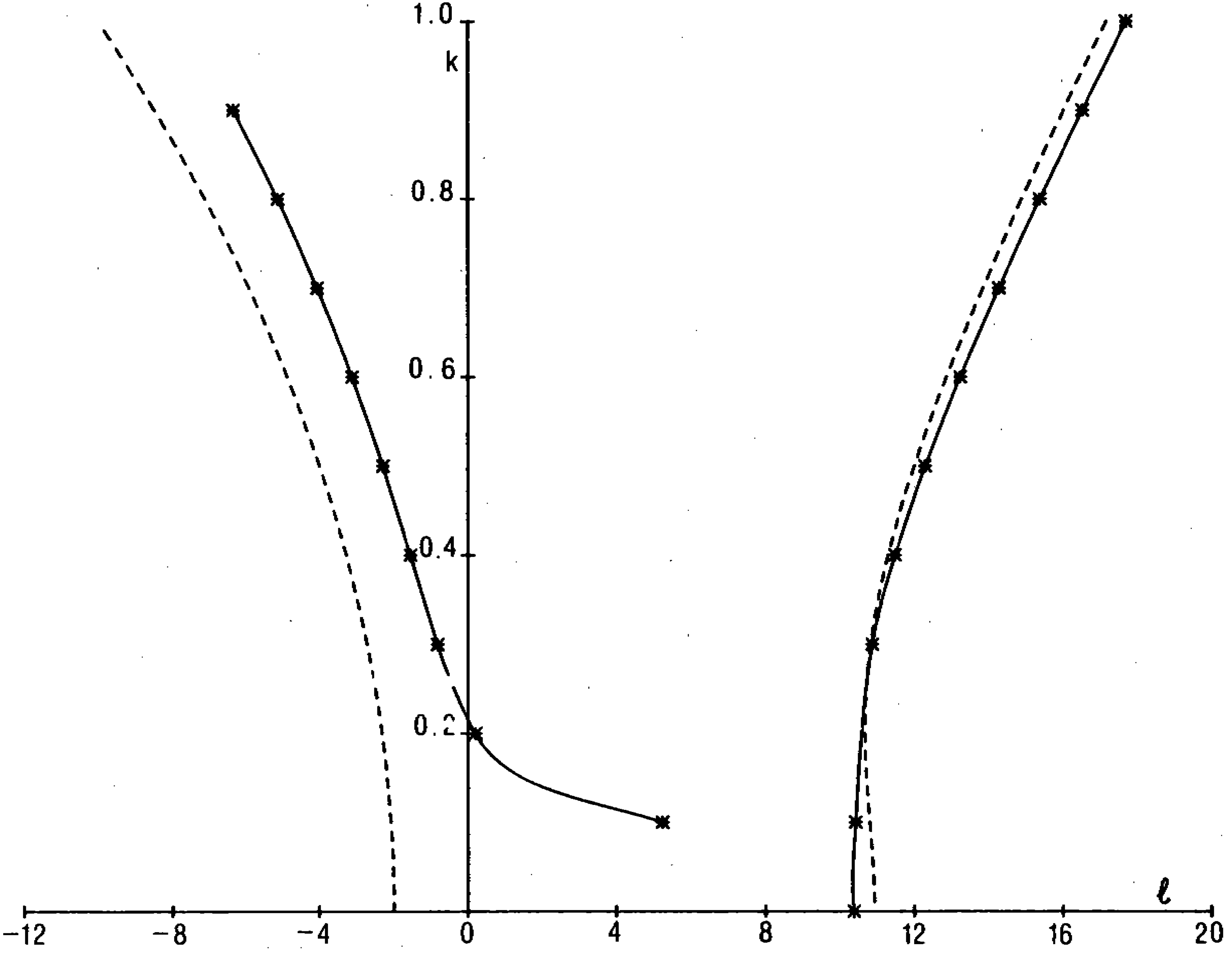


FIGURE 1

WRINKLED-FLAME CALCULATIONS REVISITED*

Helen V. McConnaughey**
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705

and

Geoffrey S. S. Ludford
Center for Applied Mathematics
Cornell University
Ithaca, NY 14853

ABSTRACT. The numerical integration of a nonlinear singular integral equation governing wrinkled plane flames is reconsidered, and the discussion extended to its counterpart for freely expanding cylindrical flames. Straightforward discretization is shown to give unsatisfactory results and a suggestion is made of how to avoid these by subtracting out the principal part of the solution near its singularity.

I. INTRODUCTION. The nonlinear singular integral equation

$$y^2(x) + ⨍_{0}^{\pi}[\cot(\tfrac{z-x}{2}) + \cot(\tfrac{z+x}{2}) - 2\cot\tfrac{z}{2}]\,y(z)\,dz = 0,\ x \in [0,\pi), \qquad (1)$$

represents a plane flame front which is distorted due to thermal expansion of the gas passing through it. The resulting profile of the flame takes the form of stationary cells which are regularly shaped and spaced.

The variable x in equation (1) is the space coordinate; the function y(x), which is odd and 2π-periodic, is proportional to the slope of the flame profile and is singular at $x = \pi$ according to:

*This research was sponsored in part by the U. S. Army under Contract No. DAAG29-80-C-0041 and Contract No. DAAG29-81-K-0127 and is based in part upon work supported by the National Science Foundation under Grant No. MCS-7927062, Mod. 2.

**Permanent address: Department of Mathematics and Statistics, Mississippi State University, Mississippi State, MS 39762.

$$y \sim 2\,\mathrm{sgn}(x-\pi)\ \ln|x-\pi| \text{ as } x\to\pi. \tag{2}$$

This result has been shown by McConnaughey, Ludford and Sivashinsky [1] in an earlier paper where equation (1) is also motivated.

Equation (1) was solved numerically in reference [1] with relation (2) replaced by $y = -2\ln(\pi - x)$ for x less than but near π. The wrinkled flame front which was calculated is shown in Figure 1.

The present report considers the more general representation of (2):

$$y = 2\,\mathrm{sgn}(x-\pi)\ \ln|\pi - x| + k \tag{3}$$

for x near π, where k is an unknown constant. It is noted that in theory, the value of k should not significantly affect the solution of (1); however, the numerical solution of a discretization of (1) subject to (3) can be sensitive to k. Application of a straightforward numerical approach to (1) and (3) for determining y(x) and k is shown to yield unacceptable results. A more promising approach is then suggested.

An equation similar to (1) describing hydrodynamic instability of freely expanding cylindrical flames is also discussed.

II. NUMERICAL SOLUTIONS. For the computation whose results are reported in reference [1] and for many of the calculations of the present study, the solution of equation (1) is approximated at discrete points $x_i = (i-1)h$, with $h = x_{N+1}/N$ and $i = 2, 3, \ldots, N-1$, by the solution of the discretized form of (1) labeled below as (4b). To obtain this discretization:

i) The upper limit of integration is replaced by $x_{N+1} = 3.13$;

ii) the trapezoidal rule is used to approximate the ordinary integrals;

iii) the principal-value integrals are treated by expanding y(z) in a Taylor series about the singular points then integrating using the trapezoidal rule or the midpoint rule; and

iv) the derivatives in the Taylor series are approximated by finite differences.

Because y(x) is an odd function, identity (4a) must hold. Finally, relation (3) is imposed at the last two mesh points. The system to be

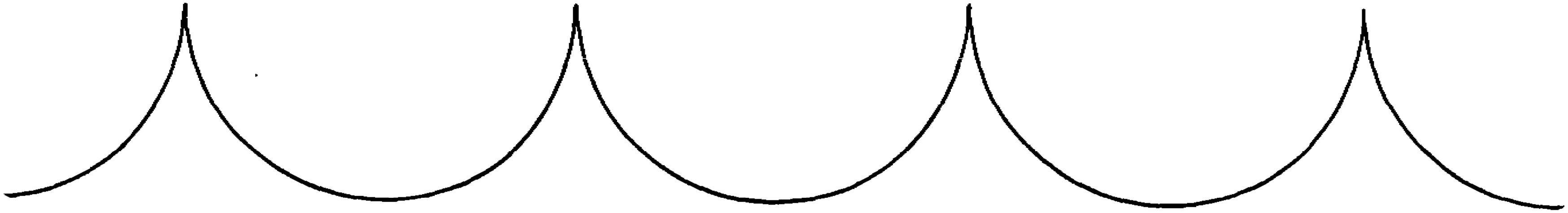


Figure 1. Shape of flame profile calculated by McConnaughey, Ludford, and Sivashinsky [1].

solved is therefore

$$y_1 = y(0) = 0; \tag{4a}$$

$$y_i^2 = y^2(x_i) = -h\{(\sum_{j=2}^{i-2} + \sum_{j=i+2}^{N})[y_j \cot(\frac{x_j - x_i}{2})] + \sum_{j=2}^{N} y_j[\cot(\frac{x_j + x_i}{2}) - 2\cot\frac{x_j}{2}] + \frac{y_{N+1}}{2}[\cot(\frac{x_{N+1} - x_i}{2}) + \cot(\frac{x_{N+1} + x_i}{2}) - 2\cot\frac{x_{N+1}}{2}]\} + (y_{i-1} - y_{i+1})(2 + \frac{h}{2}\cot\frac{h}{2}) + 2y_2, \quad i = 2, \ldots, N-1; \tag{4b}$$

$$y_i = -2\ln(\pi - x_i) + k, \quad i = N, N+1. \tag{4c}$$

The unknown constant k in expression (4c) cannot be determined analytically. If x_N and x_{N+1} could be taken sufficiently close to π, the value of k would be of little importance. The solution of (4) would vary with k only in the neighborhood of the cusp and the actual flame profile would not be substantially affected. This is difficult to accomplish in practice, however, because of the weakness of the logarithmic singularity.

For our numerical calculations, the value of k in (4c) does appreciably affect the solution of (4). Furthermore, a multitude of values for k are seen to admit a solution. When N = 32, for example, k-values between -50 and 3 give rise to a solution of (4) and many values outside of this range are expected to do likewise. However, the solutions corresponding to values of k less than -4.3 are not physically acceptable. (This critical number varies with N.)

A "reasonable" value for k may be selected by solving an appropriately modified version of system (4) which includes k as an unknown to be calculated. The extra equation needed can be obtained by imposing the discretized form of (1), in addition to relation (4c), at x_N. This adds the equation

$$y_N^2 = -h\{\sum_{j=2}^{N-2} y_j \cot(\frac{x_j - x_N}{2}) + \sum_{j=2}^{N} y_j[\cot(\frac{x_j + x_N}{2}) - 2\cot\frac{x_j}{2}]\} - y_{N+1}[\frac{h}{2}\cot(\frac{x_{N+1} + x_N}{2}) + 2 - h\cot\frac{x_{N+1}}{2}] + y_{N-1}(2 + \frac{h}{2}\cot\frac{h}{2}) + 2y_2. \tag{4d}$$

The numerical solution of (4a - d) for several values of N produces the results illustrated in Figure 2. It appears that as N increases, k may gradually tend toward some fixed value, but that limit (if it exists) remains undetermined due to the large N's required to approach it. Also, the magnitude of k becomes comparable to the value of y at x_N and x_{N+1} thereby canceling the imposed asymptotic behavior (4c) since k is negative. The numerical solution thus seems to approach the zero solution inasmuch as is possible as N increases. This trend is also demonstrated by a flattening of the profile and a decrease in the height of the cusp.

Other approximations of (1) were also considered in the present study and yielded similar results. For example, system (4a - d) was modified to include the contribution of $z \in [x_{N+1}, \pi)$ to the integral in (1). The solution did not change significantly. Also, the problem was solved when relation (4c) was imposed at the last mesh point only and a discretized form of (1) was imposed on y_i for i = 2, 3, ..., N + 1. In that case, the value of k obtained for each N was somewhat higher than that yielded by (4a - d), nevertheless the computed profile was similar (see Figure 3). In all cases investigated, the numerical solution exhibits the same behavior as N increases. This behavior is contrary to what is known about the solution from analytical considerations (relation (2)). Thus, these numerical results are not acceptable.

III. ALTERNATE APPROACH. A better way to solve equation (1) may be to introduce and solve for the smooth bounded function Y(x) given by

$$Y(x) = y(x) + 2\ln(\pi - x) - k, \quad x \in [0,\pi].$$

Then Y(x) and k satisfy the equations

$$Y(0) = 2\ln\pi - k; \quad Y(\pi) = 0;$$

$$[Y(x) - 2\ln(\pi - x) + k]^2 + \fint_0^{\pi} [\cot(\frac{z-x}{2}) + \cot(\frac{z+x}{2}) - 2\cot\frac{z}{2}][Y(z) -$$

$$2\ln(\pi - z) + k]dz = 0, \quad 0 < x < \pi; \tag{5}$$

$$k^2 + \int_0^{\pi} [2\cot(\frac{z-\pi}{2}) - 2\cot\frac{z}{2} - \frac{4}{z-\pi}][Y(z) - 2\ln(\pi - z) + k]dz -$$

$$4k\ln\pi + \fint_0^{\pi} \frac{4}{z-\pi} Y(z)dz + A + 4\int_b^{\infty} (\frac{1}{s-1} - \frac{2}{s} + \frac{1}{s+1})\ln s\, ds = 0,$$

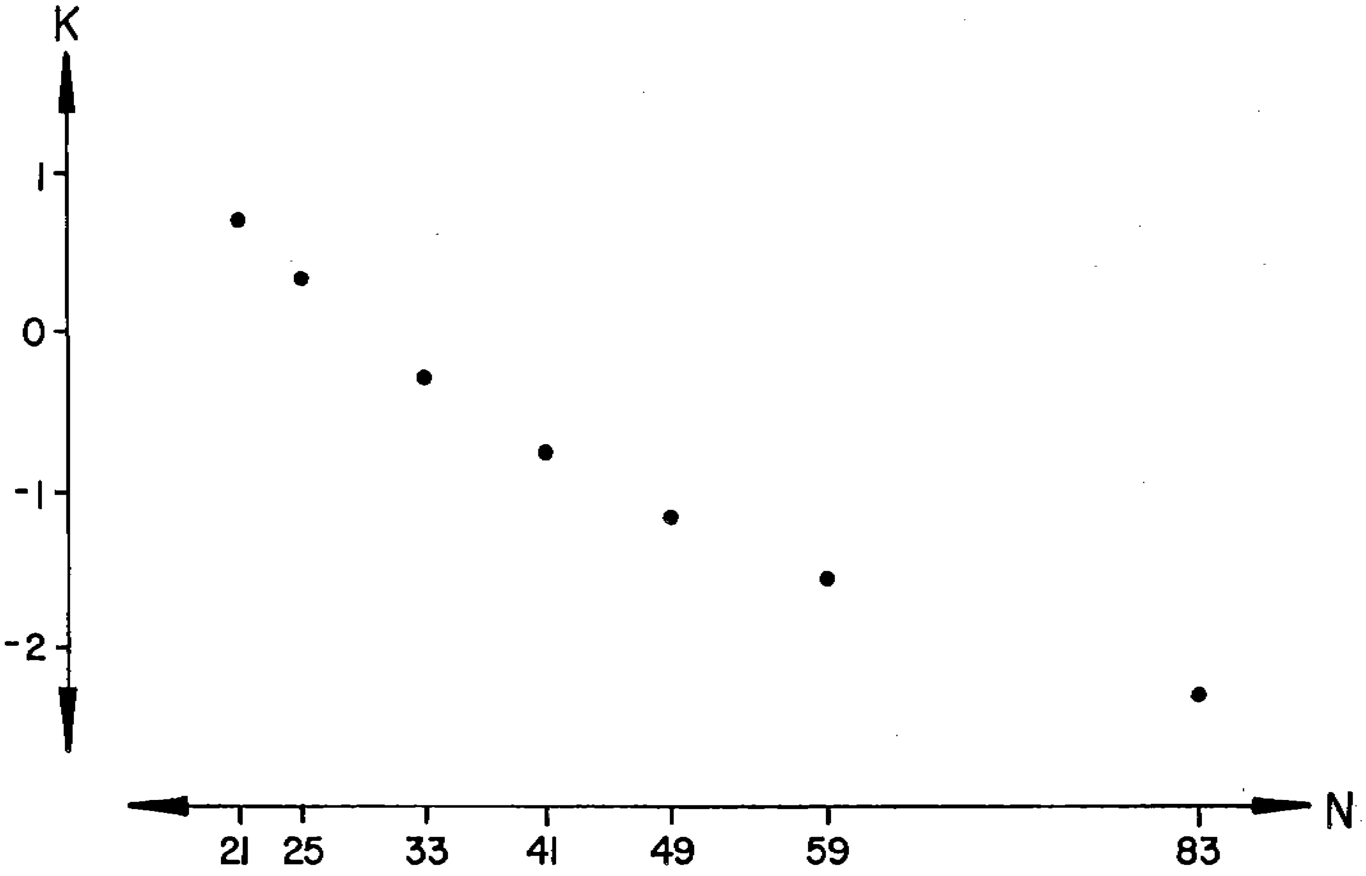


Figure 2a. Values of k calculated for different N's. For k = 27, k lies between .1 and .2 as compared to the zero value assumed in reference [1].

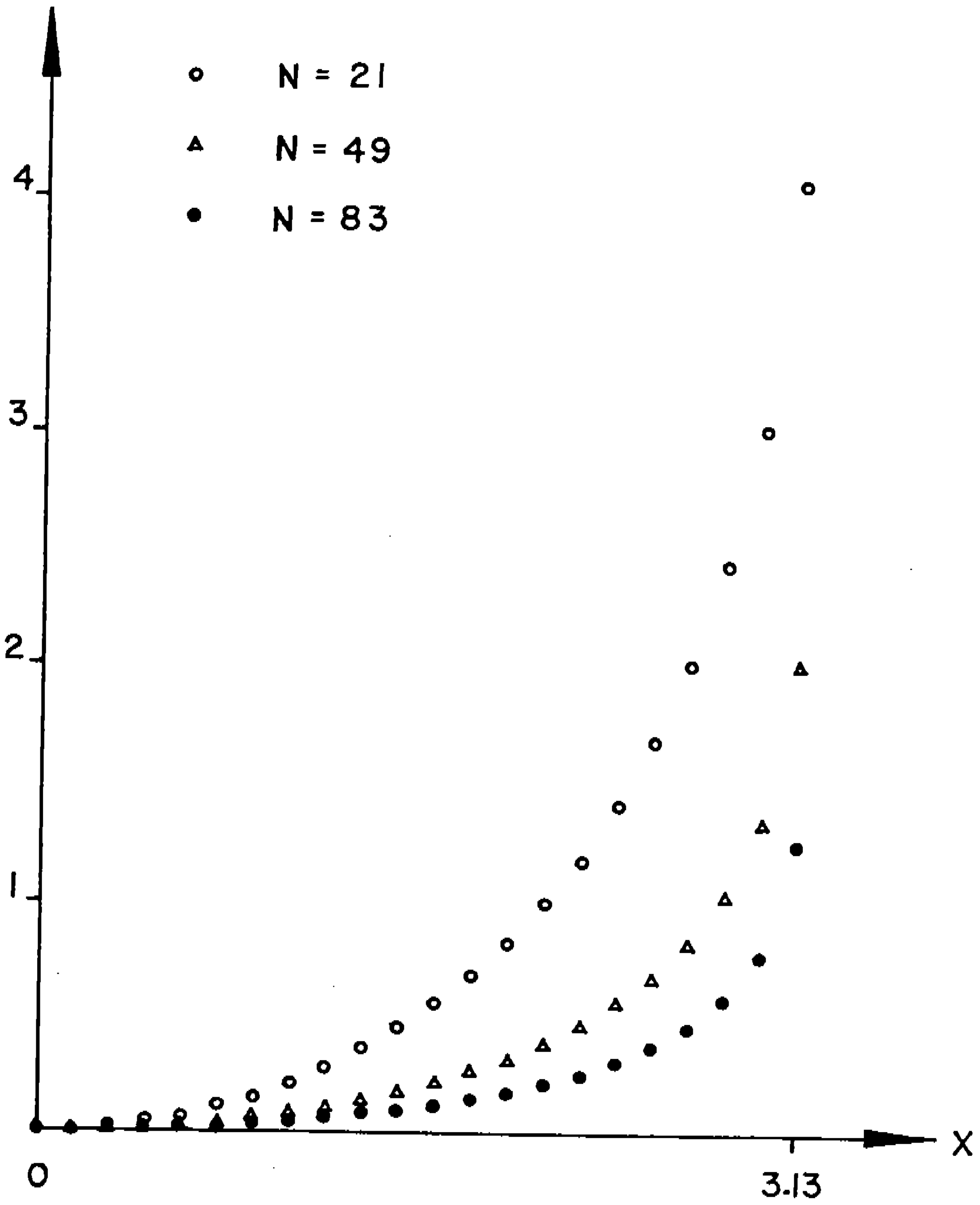


Figure 2b. Flame profile calculated for three values of N.

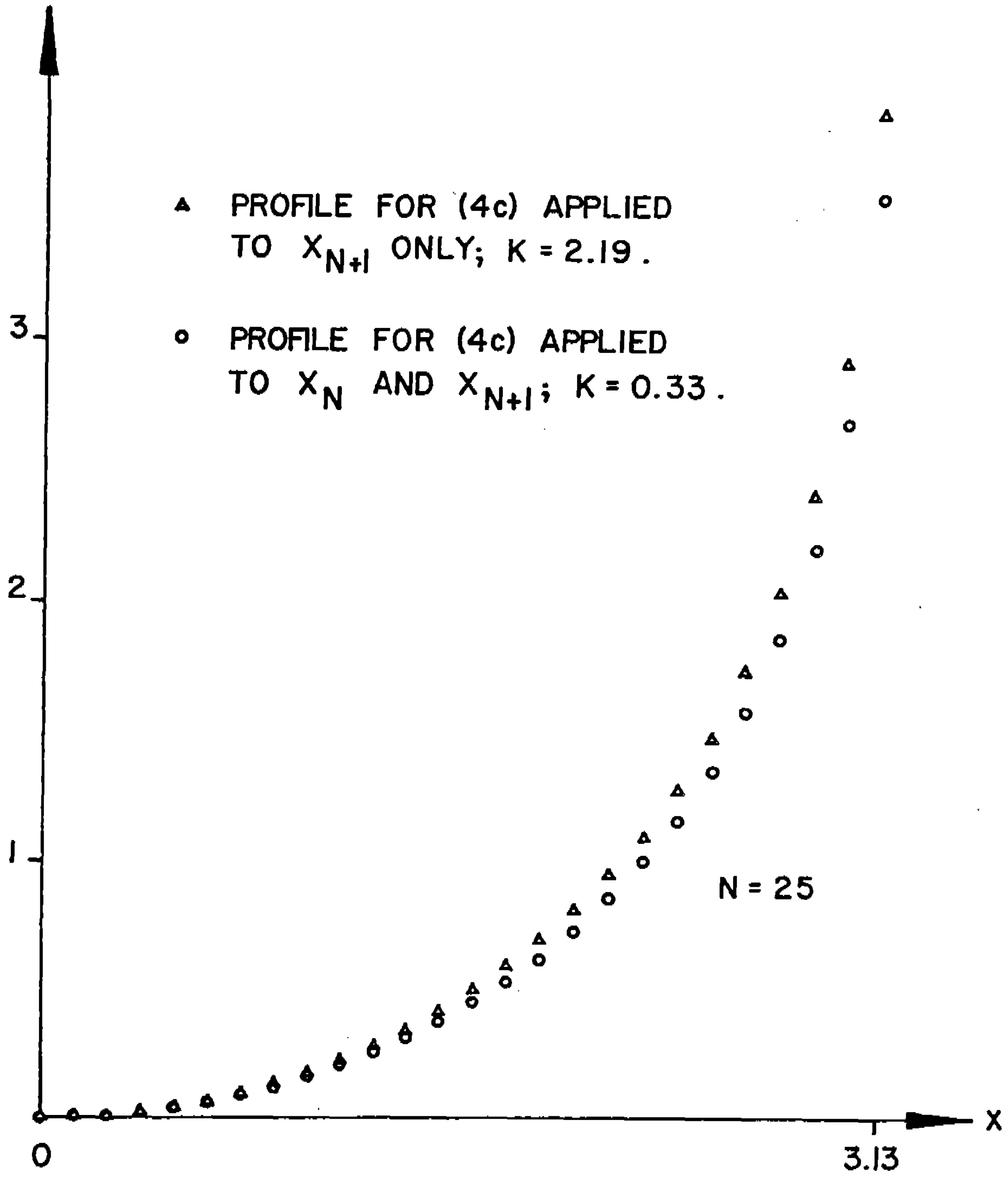


Figure 3. Flame profile calculated for N = 25 from the solution of system (4a-d) and from solution of the similar system in which (4c) is imposed at x_{N+1} only. The value of k obtained in the first case is .33 and is 2.19 in the second case.

where $b > 1$ and

$$A = 4[\ln^2\pi - \ln^2 b + \fint_0^b (\frac{1}{s-1} + \frac{1}{s+1})\ln s\, ds].$$

The solution of this system is presently under numerical investigation. For x near π, equation (5) is rewritten as

$$[Y(x) + k]^2 + 4[\ln x + \ln(2\pi - x) - 2\ln\pi - Y(x)]\ \ln(\pi - x) +$$

$$\int_0^\pi I(x,z)[Y(z) - 2\ln(\pi - z) + k]dz - 2k[\ln x + \ln(2\pi - x)] +$$

$$\fint_0^\pi (\frac{2}{z-x} - \frac{2}{2\pi - z - x})Y(z)dz + A + 4\int_b^{\pi/(\pi-x)} (\frac{1}{s-1} - \frac{2}{s} + \frac{1}{s+1})\ln s\, ds = 0,$$

where $I = \cot(\frac{z-x}{2}) - \frac{2}{z-x} + \cot(\frac{z+x}{2}) + \frac{2}{2\pi - z - x} - 2\cot\frac{z}{2}$. In this form, large terms in (5) which are potentially troublesome have been subtracted then added in such a way as to remove possible difficulties. It is assumed that $Y(x) \to 0$ faster than $1/\ln(\pi - x)$ as $x \to \pi$.

IV. CYLINDRICAL FLAMES. The equation analogous to (1) for the slope of a perturbed, freely expanding cylindrical flame is

$$y^2(\theta) + \fint_0^\pi [\cot(\frac{\phi-\theta}{2}) + \cot(\frac{\phi+\theta}{2}) - 2\cot\frac{\phi}{2}]\,y(\phi)d\phi + c\int_0^\theta y(\phi)d\phi = 0, \qquad (6)$$

where θ is a suitably scaled polar angle and c is a physical parameter for which a typical value is near 16. The singular behavior of $y(\theta)$ at $\theta = \pi$ is equivalent to that given in (2). Note that for $c = 0$, equation (6) is the same as equation (1).

The numerical approach that was applied to (1) has also been applied to (6). However, only for small values of c has a solution been found. For realistic values of c, no meaningful results have yet been obtained.

V. CONCLUSION. The value of the constant k which makes the asymptotic statement

$$y \sim 2\ \mathrm{sgn}(x - \pi)\ \ln|x - \pi| + k \text{ as } x \to \pi$$

consistent with equation (1) or equation (6) is seen to have importance in numerical calculations of the solution of those equations. Attempts to find k by a simple and straightforward modification of the numerical approach in reference [1] (where k is assumed to be zero) are shown to fail. Another approach, which changes equation (1) to a problem whose solution has no singularities, is suggested but not pursued here.

REFERENCES

[1] H. V. McConnaughey, G. S. S. Ludford, and G. I. Sivashinsky (1983). A calculation of wrinkled flames. Combustion Science and Technology 33, 103.

NEAR CHAPMAN-JOUGET DETONATIONS*

D.S. Stewart
Department of Theoretical and Applied Mechanics
University of Illinois, Urbana-Champaign, IL 61801

G.S.S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

ABSTRACT. The moving boundary problem that arises from a certain model of one-dimensional detonation waves is considered. Burgers equations for a function f have to be solved on the two sides of a discontinuity, at which f and the jump in its derivative (corresponding to the exothermic reaction) are prescribed. The steady solution, a summary of its linear stability characteristics, and some preliminary numerical results are presented.

I. INTRODUCTION. We consider a simple model of plane detonation waves in which all the reaction is supposed to occur at a single location and at a prescribed (ignition) temperature. For combustion waves that travel at finite Mach numbers, the model implies very fast kinetics. A complete presentation of the flame-sheet theory on which the present discussion rests can be found in Stewart and Ludford [1].

Substantial simplification of the mathematical problem occurs when the heat released by the combustion is small, as was seen last year when we presented the corresponding (but fundamentally different) results for near Chapman-Jouget deflagrations [2]. Small heat release leads to a gasdynamic state close to the constant state ahead of the wave. Thus, the temperature and pressure are, respectively,

$$1 + \beta^{\frac{1}{2}}(\gamma - 1)f + O(\beta) \ , \quad 1 + \beta^{\frac{1}{2}}\gamma f + O(\beta) \ , \tag{1}$$

and the velocity of the detonation wave is

$$-1 - \beta^{\frac{1}{2}}c \ . \tag{2}$$

*Work supported by the U. S. Army Research Office.

Here γ is the ratio of specific heats, β is the (dimensionless) heat released, c measures the closeness of the wave velocity to the undisturbed sound speed (-1), and f represents the variation of the gasdynamics state from the undistrubed state. Following [2] we will use η for distance in a frame moving with the wave and T for time.

The function $f(\eta,T)$ satisfies

$$f_T + \frac{\gamma+1}{2}\left(k(T) - f\right)f_\eta = \frac{\gamma}{2} f_{\eta\eta} , \tag{3}$$

$$f(0,T)=f_*,\ f_\eta(0+,T) - f_\eta(0-,T) = -Y_o/\gamma \equiv -(\gamma+1)\alpha^2/2\gamma, \tag{4}$$

$$f(-\infty,T) = 0 \ ,\quad f(+\infty,T) < \infty \ , \tag{5}$$

where

$$k = 2c/(\gamma+1) \tag{6}$$

will also be called the velocity of the wave. Here $\eta = 0$ is the location of the flame sheet (where the temperature is constant), f_* is a reactivity parameter fixing the flame temperature, and Y_o is the mass fraction of the deficient reactant. If c(T) were known, only four boundary conditions would be required to determine f . However, there are five because the condition (4a) counts twice; so that c is determined along with f .

II. STEADY SOLUTIONS AND THEIR STABILITY.

From the Rankine-Hugoniot jump conditions it can be shown that there is a minimum steady detonation velocity, the corresponding wave being know as a Chapman-Jouget detonation, here given by $k = \alpha$. for every $k \geq \alpha$ and $f_* \in [f_-,2k)$, there is a steady solution

$$f = \begin{cases} 2kf_*e^{\zeta_-}/(2k - f_* + f_*e^{\zeta_-}) \\ \left[f_-(f_+-f_*)+f_+(f_*-f_-)e^{\zeta_+}\right]/\left[(f_+-f_*)+(f_*-f_-)e^{\zeta_+}\right] \end{cases} \text{for } \eta \lessgtr 0, \tag{7}$$

where

$$f_{\pm}=k\pm\sqrt{k^2-\alpha^2}\ ,\quad \zeta_-=(\gamma+1)k\eta/\gamma\ ,\quad \zeta_+=(\gamma+1)(f_+-f_-)\eta/2\gamma\ . \qquad (8)$$

The solutions corresponding to $f_* = f_-$ and $f_* \to 2k$ have special significance in detonation theory. For $f_* = f_-$, the weak detonation is obtained, the velocity k being uniquely determined by the reactivity parameter f_* . As $f_* \to 2k$, the ZND detonation is obtained, i.e. a shock with velocity k followed by a deflagration adjusting f from $2k$ to f_+ .

The linear stability of a steady solution can be examined by taking the infinitesimal perturbations of f and c (or k) proportional to $\exp(\lambda T)$. Stability, which is determined by the sign of $\mathrm{Re}(\lambda)$, is found to depend on the value of f_* . (When this mode analysis fails to find instability, solution of the corresponding initial-value problem by Laplace transformation confirms that there is indeed stability, though for Chapman-Jouget detonations it is neutral.) The steady detonation is found to be

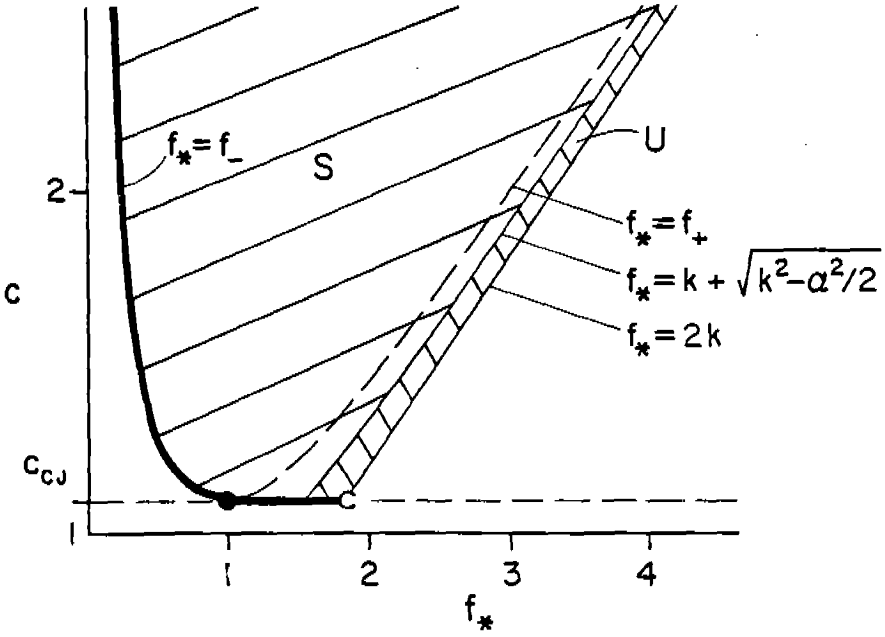


Fig. 1. The f_* , c-plane for $\gamma = 1.4$, $Y_o = 1$, showing the region where the steady solution exists and its stability. S = stable , U = unstable ; the CJ-solutions are marginally stable.

$$\text{stable for} \quad f_- \leq f_* < k + \sqrt{k^2 - \alpha^2/2} \; , \tag{9}$$

$$\text{unstable for} \quad k + \sqrt{k^2 - \alpha^2/2} \leq f_* < 2k \; . \tag{10}$$

Figure 1 shows the region of the f_* , k-plane where the steady solution exists and its stability.

Note that

$$f_* = k + \sqrt{k^2 - \alpha^2/2} \tag{11}$$

is not a conventional stability boundary: the single (real) eigenvalue by which instability is characterized tends to $+\infty$ as the boundary is approached.

III. NUMERICAL RESULTS. We now discuss the numerical solution of the problem (3-5) under various initial conditions. A description of the implicit numerical scheme used can be found in [1]; a spatial step of 0.1, a time step of 0.05, and boundaries at $\eta=\pm 10$ were used for most of the calculations shown.

A numerical test of the theoretical stability boundary (11) leads to Figure 2. The initial data correspond to the steady solution for $\gamma = 1.4$, $Y_O = 1$ (used in all the computations shown), and $c = 2$; the disturbance corresponds to the discretization error. The $c(T)$ profile was computed along with $f(\eta,T)$ for various values of f_* , so as to determine the numerical stability boundary. (The criterion for numerical stability was that f and c should tend to their steady-state values in the computation time allotted.)

In Figure 2 the theoretical region of stability is hatched. Figure 2a shows the values of c at $T = 2.5$ for different values of f_* ; the computed stability limit seems to occur at $f_* = 3.25$, while the theoretical value (11) is $f_* = 3.20$. Typically the disturbance of c (due to discretization error) has a maximum that grows as f_* is increased towards the stability limit; for larger values of f_* , the disturbance does not die out, i.e. no steady state is achieved. Details of the c,T-profiles in the hatched region of Figure 2b are shown in Figure 3.

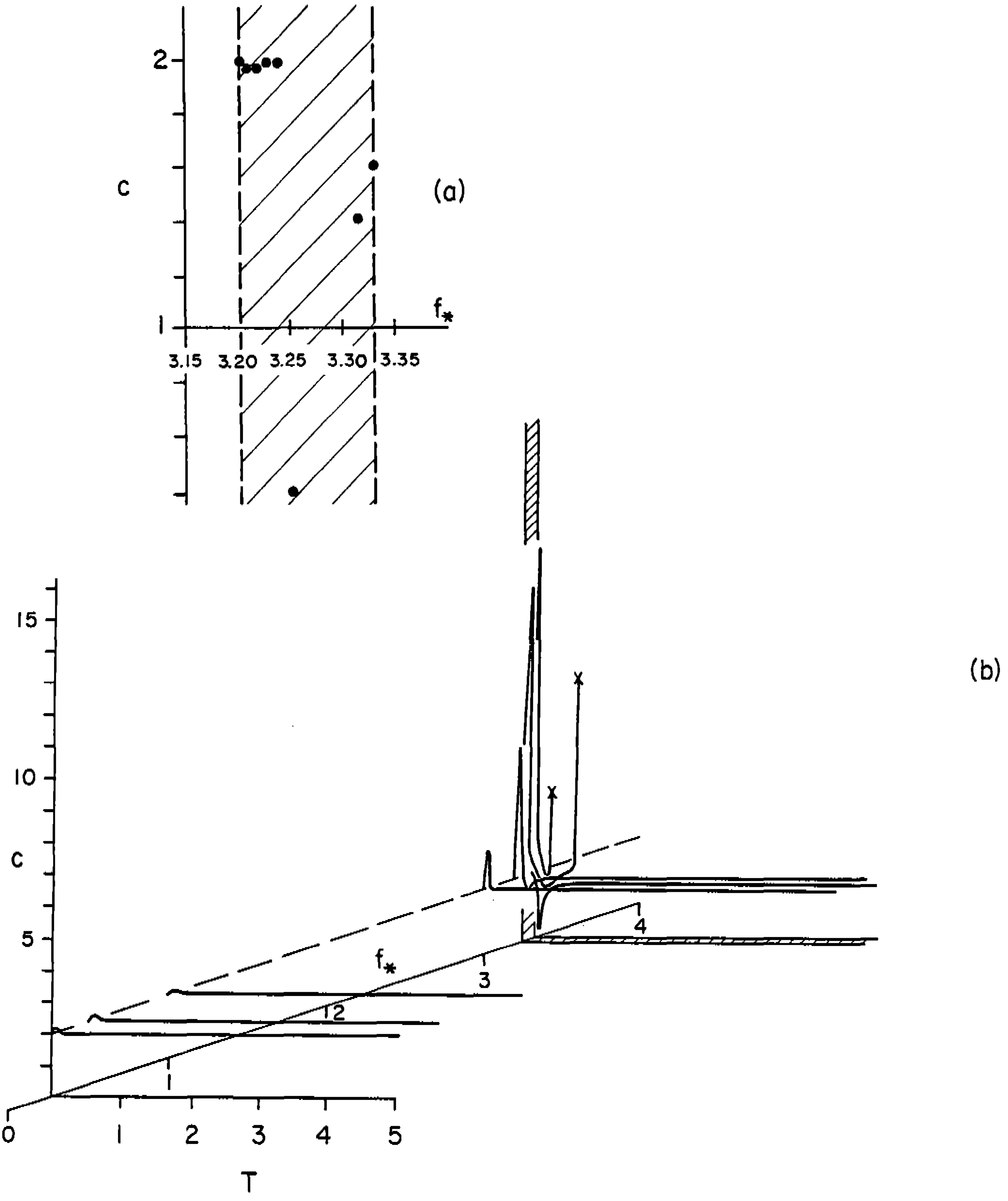


Fig. 2. Numerical test of the stability of the steady solutions.

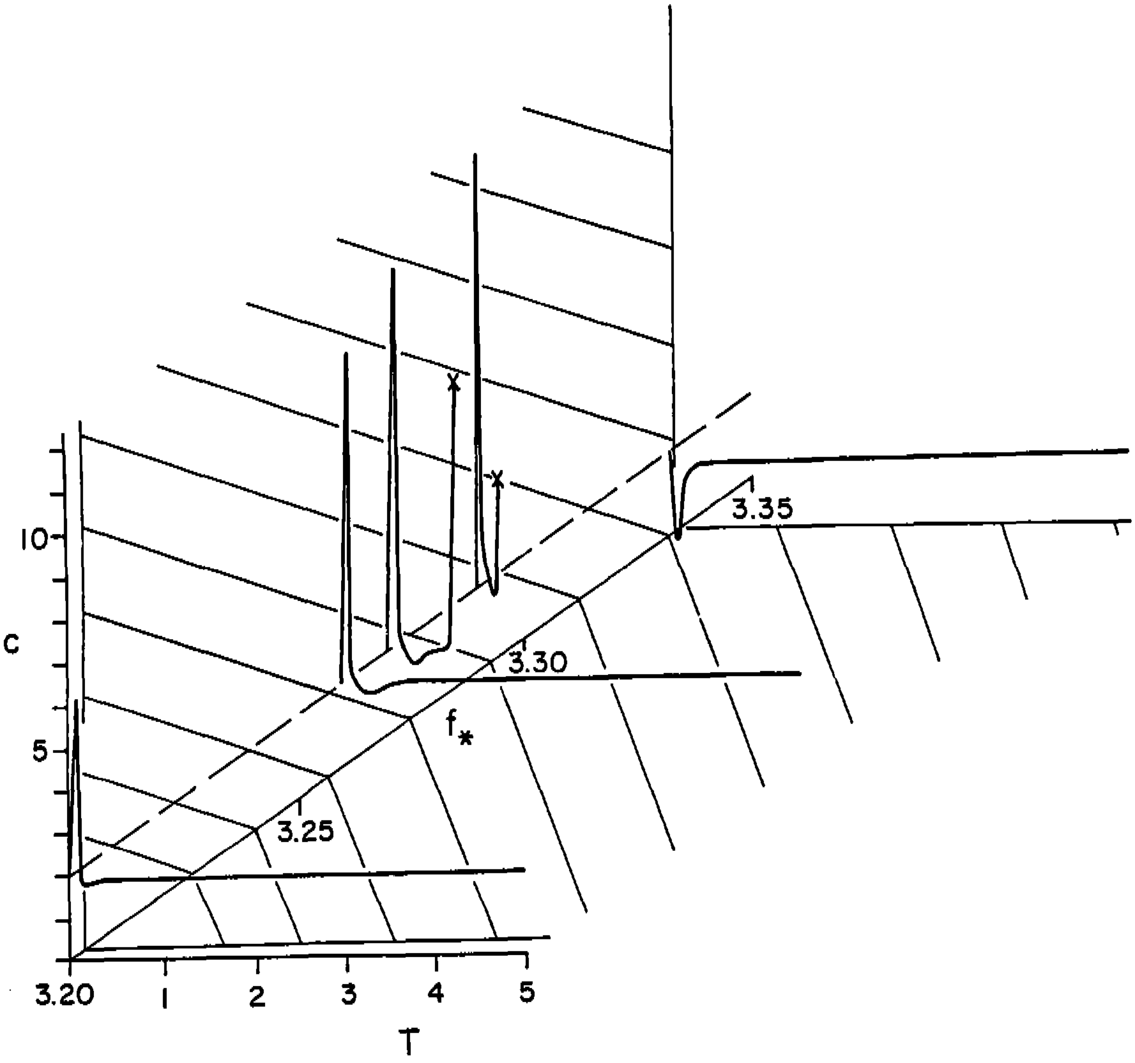


Fig. 3. Blow-up of Figure 2(b).

Figure 4 shows the response of the stable detonation $f_* = 1$, $c = 1.2$ to a larger disturbance in the burnt mixture $(n > 0)$. The dotted line is the steady solution and, since f is proportional to the pressure, the initial data shown correspond to a rarefactive disturbance.

Figure 5 shows the response of the unstable detonation $f_* = 1.8$, $c = c_{CJ} = 1.09545$ to a similar rarefaction. Figure 6 shows the corresponding response for the detonation $f_* = 1.5$, $c = c_{CJ}$, which is neutrally stable but close to the theoretical stability limit. The numerical scheme failed at $T = 0.1992$, but a rudimentary oscillation in c is obtained, with a small but nontrivial change in f .

Figure 7 shows the same detonation as in Figure 6, but subjected to a stronger rarefaction. (These results were obtained by an explicit scheme, with spatial step 0.2 and time step 0.025.) Almost three cycles of an apparently periodic oscillation were observed before the scheme failed (at X).

IV. CONCLUDING REMARKS. The numerical calculations correlate reasonably well with the theoretical stability results. Certainly the large region of stability is confirmed, and inaccuracy in locating its boundary could well be due to the discretization error.

Of greatest interest is that our results suggest the possibility of relaxation oscillations for our model. Such oscillations have been observed in numerical simulations of detonations using a system of first-order convective-reactive equations [3], and have been dubbed "galloping detonations."

We are currently trying to improve our numerical scheme to overcome the difficulties that are encountered when c is large. This should make the evidence for relaxation oscillations more convincing.

REFERENCES

[1] D.S. Stewart and G.S.S. Ludford. The acceleration of fast deflagration waves. Z. angew. Math. Mech. (in press).

[2] D.S. Stewart and G.S.S. Ludford. Evolution of near Chapman-Jouget deflagrations. Transactions of the 28th Conference of Army Mathematicians, pp. 133-142, ARO-Report 83-1, 1983.

[3] W. Fickett and W.C. Davis. Detonation, p. 276. Berkeley: University of California Press, 1979.

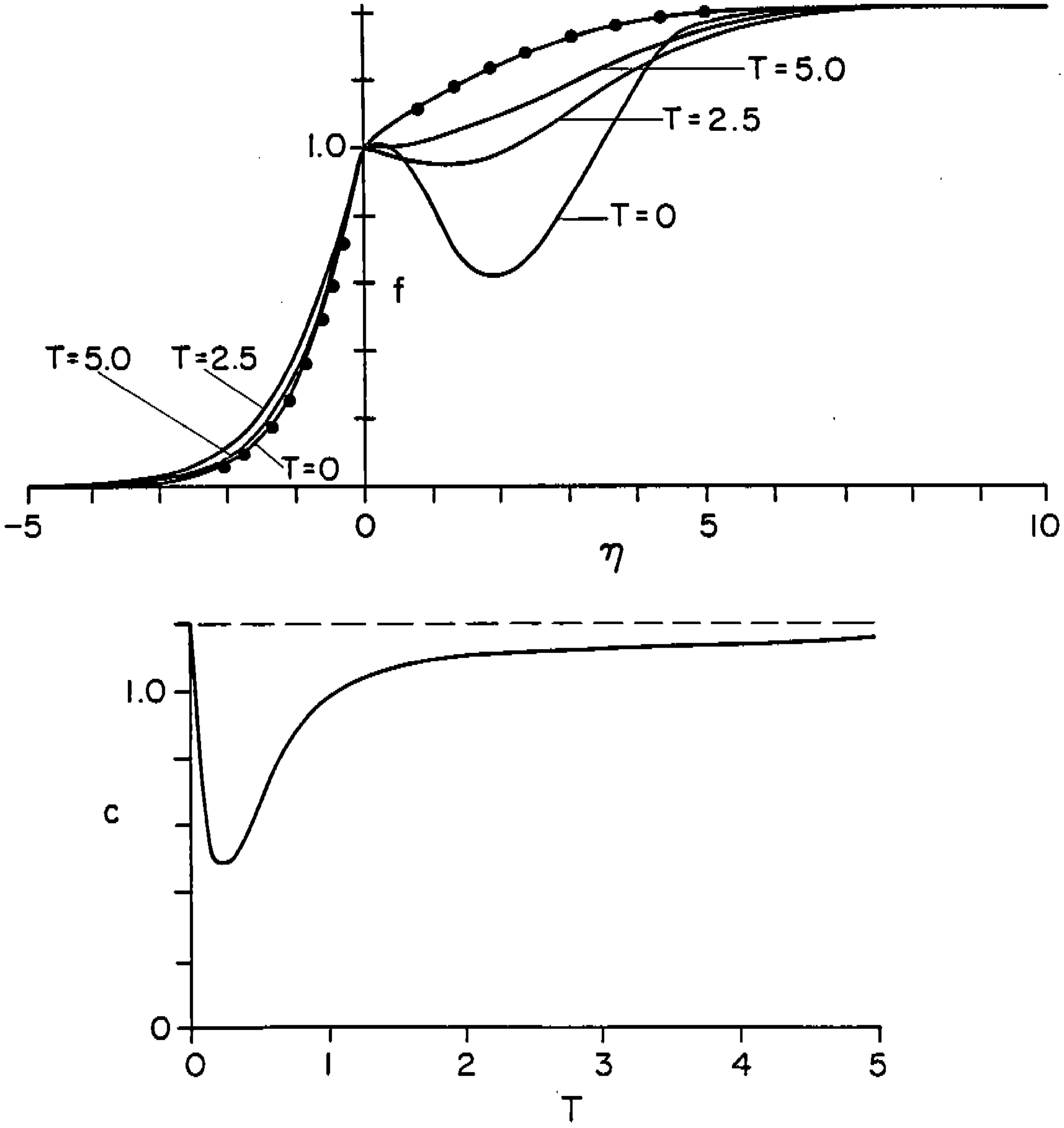


Fig. 4. Response of a stable detonation to a rarefaction applied in the burnt region. The curve ●—●—● is the undisturbed steady profile.

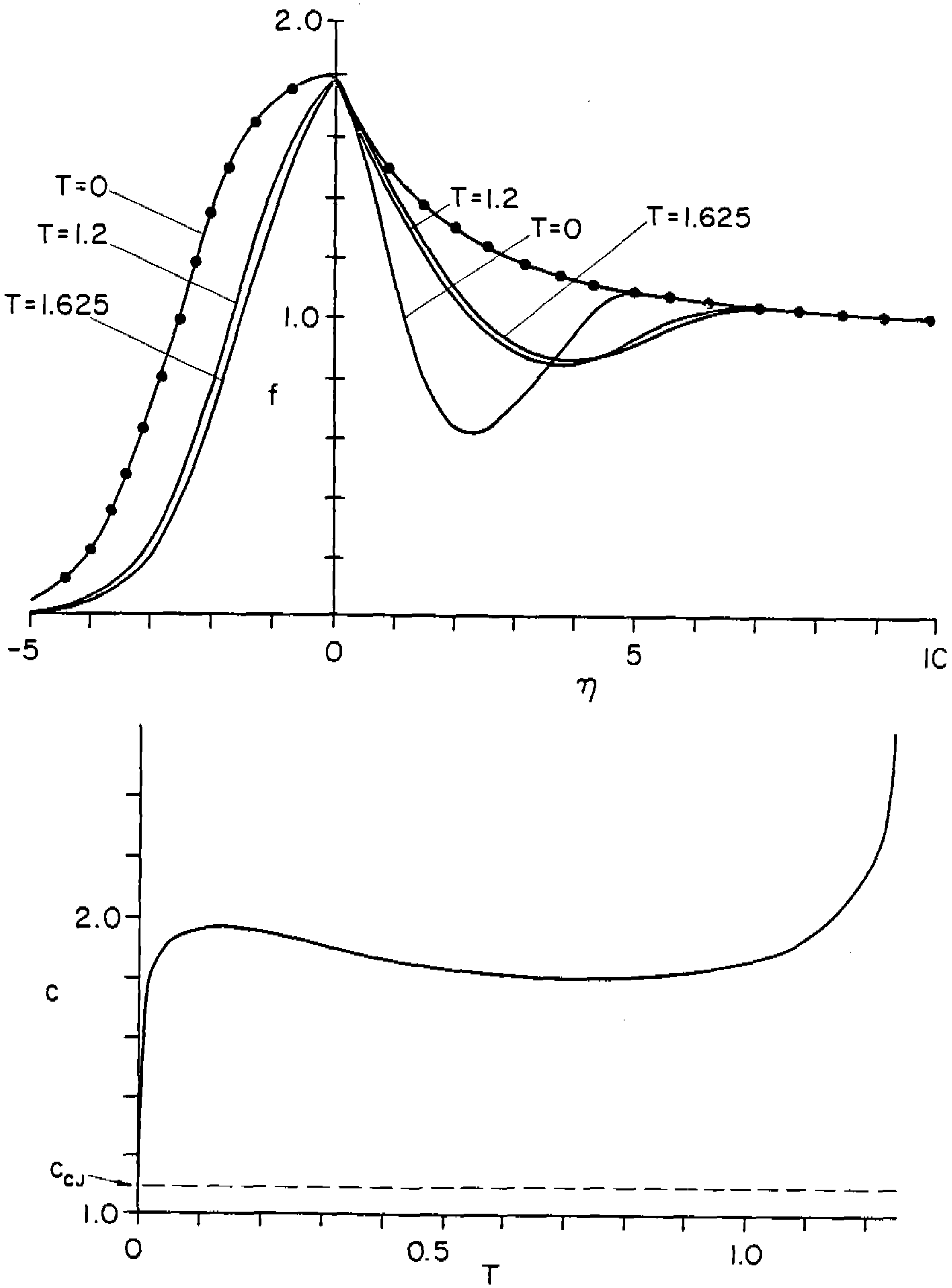


Fig. 5. Response of an unstable detonation to a rarefaction in the burnt region. The curve —●—●—●— is the undisturbed steady profile.

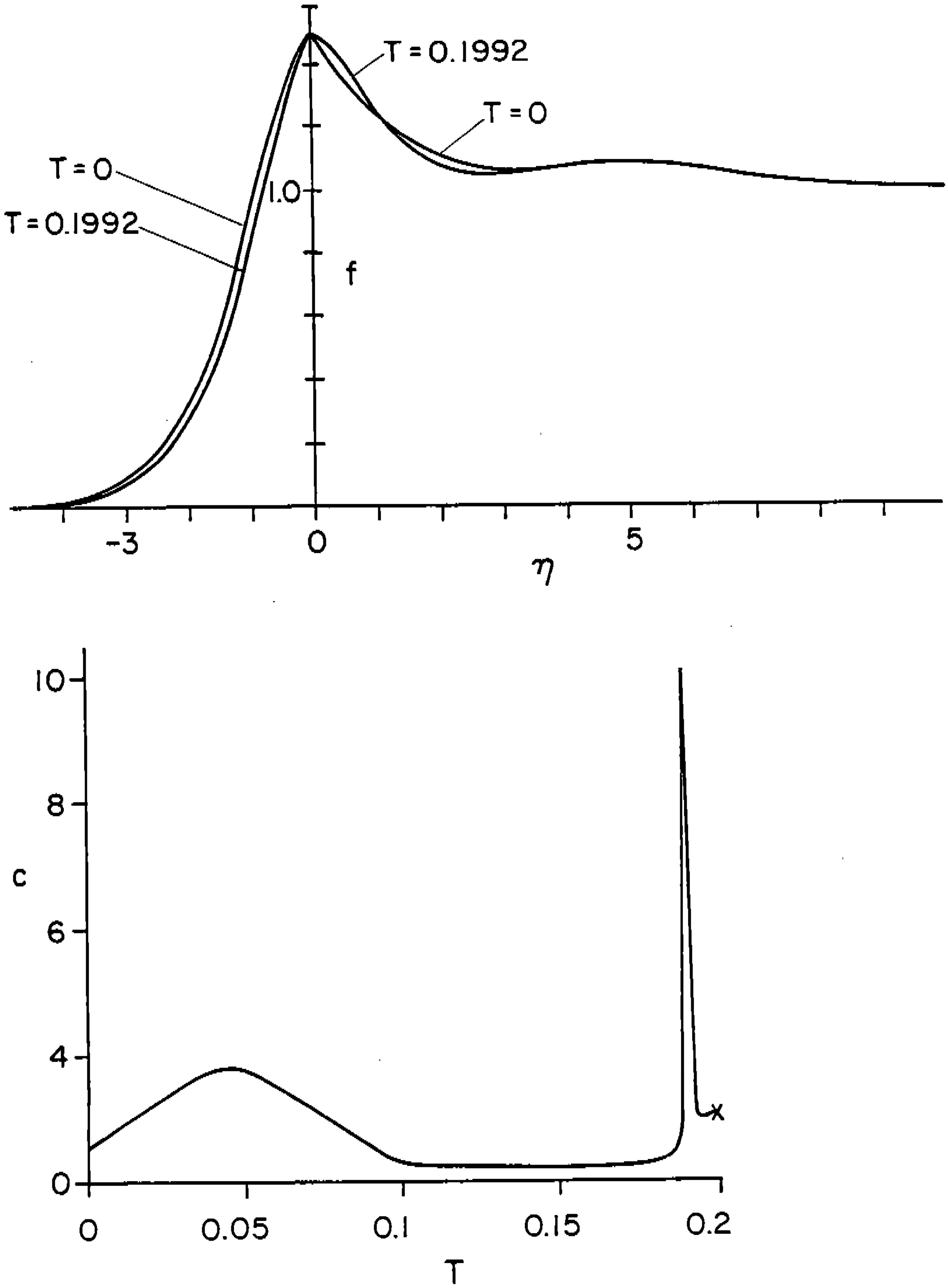


Fig. 6. Response of a neutrally stable CJ-detonation to a rarefaction in the burnt region. Breakdown at T = 0.1992 is not shown by f-profile.

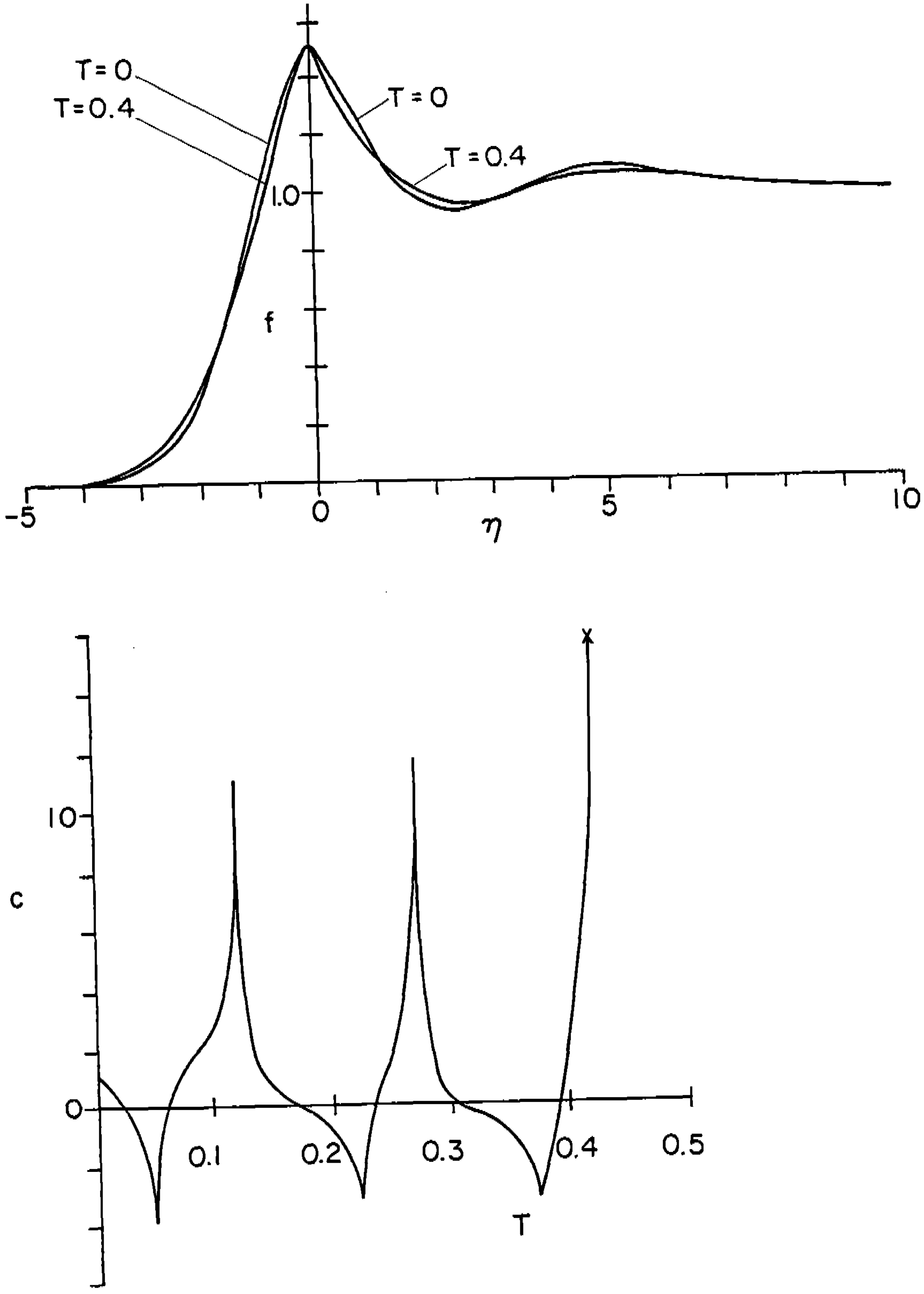


Fig. 7. Response of CJ-detonation in Figure 6 to a slightly stronger rarefaction. At breakdown the f-profile is indistinquishable from that shown for T = 0.4 .

ARE DETONATIONS STEADY?*

A.A. Oyediran and G.S.S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, New York 14853

ABSTRACT. A generic problem in the transition from deflagration to detonation is the overtaking of a steady deflagration wave by a steady shock wave. Such a collision produces a detonation wave (as well as a contact discontinuity, a back shock, and sometimes a rarefaction wave). Work reported at the 27th Conference of Army Mathematicians showed that, for small heat release in the deflagration, the detonation wave cannot be steady. Here we remove the restriction to small heat release and show that the detonation wave can never be steady.

I. INTRODUCTION. In a paper presented at the 27th Conference of Army Mathematicians, Ludford and Stewart (1) considered the shock-induced transition from deflagration to detonation illustrated in Figures 1 and 2. In particular, they showed that, for small heat release, the resulting detonation wave cannot be steady. The object of this paper is to remove the restriction to small heat release: the detonation wave that results from a steady shock overtaking a steady deflagration wave can never be steady, according to the ignition-temperature theory of deflagrations and detonations developed in (2) and (3).

II. RESULTS FROM THE THEORY. The reactivity of the fresh mixture ahead of both deflagration and detonation is characterized by an ignition temperature T_*. Stewart & Ludford (2) then show that a structure exists for the steady deflagration only if the ignition temperature lies in a certain interval

$$\overset{\smile}{T}_*(\beta) \le T_* < \overset{\frown}{T}_*(\beta) = 1 + \beta \; ; \qquad (1)$$

*Work supported by the U.S. Army Research Office

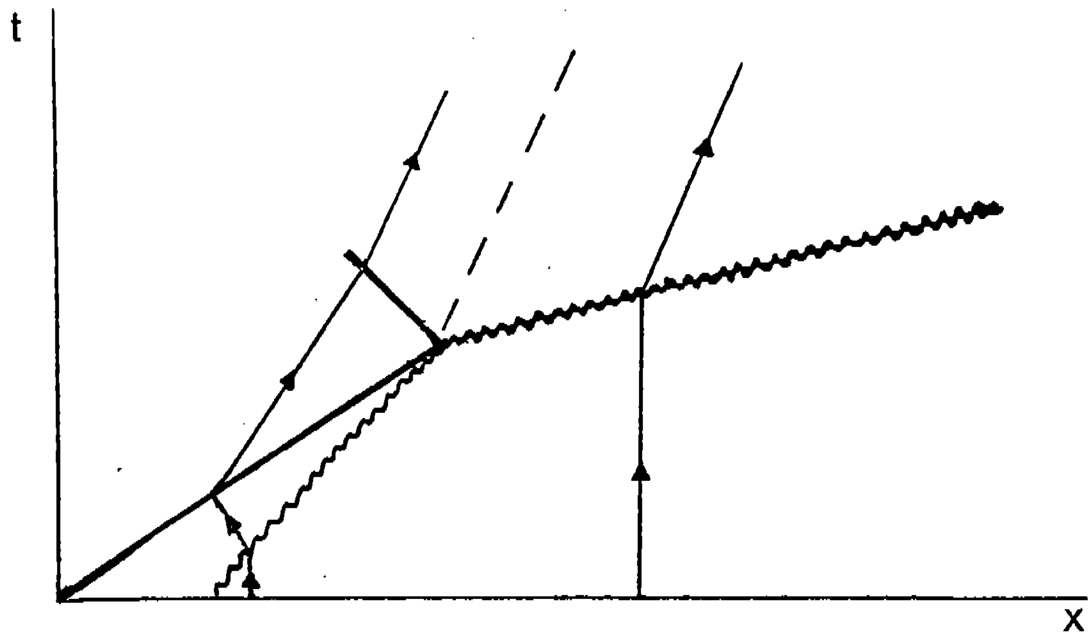


Figure 1. Outcome of a sufficiently strong steady shock ━━━ overtaking a steady deflagration ∿∿∿. The lines ∿∿∿ and — — represent a detonation and contact discontiniuity, respectively.

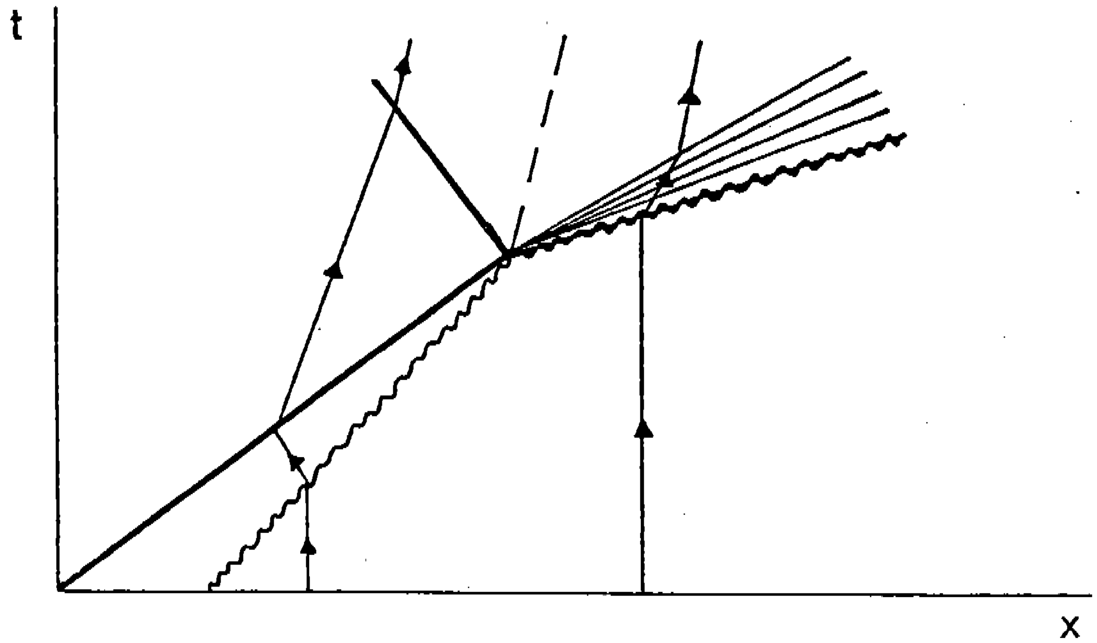


Figure 2. Modification of Figure 1 for weaker shock strengths. The detonation is now Chapman-Jouget with a centered rarefaction lying immediately behind.

here β is the heat released by the reaction and the temperature of the fresh mixture has been used to non-dimensionalize. The right limit is the so-called adiabatic flame temperature, when no heat is used to increase the kinetic energy; the left limit, which will not be used in our discussion, must be obtained by numerical integration.

Given the Mach number M of a detonation, it is well known (4) that the reaction rate must lie within certain limits. For the ignition-temperature theory, Lu & Ludford (3) show that this leads to the requirement

$$\check{T}_*(\beta,M) \leq T_* \leq \hat{T}_*(\beta,M) \quad , \tag{2}$$

where

$$\check{T}_*=1+\beta+\frac{\gamma-1}{2}M^2(1-v^2) \text{ with } v=\frac{1+\gamma M^2+\sqrt{(M^2-1)^2-2\beta(\gamma+1)M^2}}{(\gamma+1)M^2} \tag{3}$$

but $\hat{T}_*$ must be obtained by numerical integration (we shall not need it). Here γ is the specific-heat ratio and we find

$$(1+\gamma M^2)/(\gamma+1)M^2 < v < 1 \; . \tag{4}$$

III. APPLICATION TO THE PRESENT PROBLEM. The Mach number of the detonation in either figure is determined by the strength of the shock wave. It cannot be smaller than the Chapman-Jouget value, i.e,

$$M_{CJ} \leq M < \infty \quad \text{with} \quad M_{CJ}^2 = 1+(\gamma+1)\beta+\sqrt{(1+(\gamma+1)\beta)^2-1} \; . \tag{5}$$

When the shock is sufficiently strong, as in Figure 1, the Mach number is greater than M_{CJ} ; otherwise it equals M_{CJ} , as in Figure 2. We shall now show that, whatever the Mach number, the intervals (1) and (2) do not overlap; specifically

$$\check{T}_*(\beta,M) > \hat{T}_*(\beta) \quad \text{for all} \quad M \,\varepsilon\, (M_{CJ},\infty) \; , \tag{6}$$

which is clearly ensured if the following inequalities hold:

$$\check{T}_*(\beta,M_{CJ}) > \check{T}_*(\beta,\infty) > \hat{T}_*(\beta) \; , \tag{7}$$

$$d\check{T}_*/dM < 0 \quad \text{for all} \quad M \,\varepsilon\, (M_{CJ},\infty) \; . \tag{8}$$

Proof of the inequalities (7) follows directly from the definition (3); thus,

$$\check{T}_*(\beta, M_{CJ}) = 1 + \gamma\beta + (\gamma-1)\{\sqrt{(1 + (\gamma+1)\beta)^2-1} - \beta\}/(\gamma+1)$$

$$> 1 + \gamma\beta = \check{T}_*(\beta,\infty) > 1 + \beta = \hat{T}_*(\beta) \ .$$

The definition also gives

$$d\check{T}_*/dM = (\gamma-1)M(1 - v^2 - 2vM^2dv/dM^2)$$

$$= -(\gamma-1)M(1-v)[(1+\gamma M^2)+(1-\gamma M^2)v]/2[(\gamma+1)M^2v-(1+\gamma M^2)] \ ,$$

where we have used the quadratic equation satisfied by v rather than its solution (3b). The bounds (4) on v now give the inequality (8).

REFERENCES

(1) G.S.S. Ludford & D.S. Stewart. Deflagration to detonation transition. *Transactions of the 27th Conference of Army Mathematicians*, pp. 563-572, ARO-Report 82-1, 1982.

(2) D.S. Stewart & G.S.S. Ludford. Fast deflagration waves. *J. Meć.* (in press).

(3) G.C. Lu & G.S.S. Ludford. Asymptotic analysis of plane steady detonations. *SIAM J. Appl. Math.* *42*, pp. 625-635 (1982).

(4) F.A. Williams. *Combustion Theory*, p. 144. Reading (Mass.): Addison-Wesley Publishing Co., 1965.

THERMAL AND TRANSFORMATION STRESSES IN HOLLOW TUBES DURING THE QUENCHING PROCESS

J. D. Vasilakis
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. During the heat treatment of components, the transient thermal stresses can be very high. This is especially true if a severe quench is required such as the quenching of steel gun tubes for the development of a martensitic grain structure. In addition to the large transient thermal stresses, severe transformation stresses also exist due to the structural volume change involved. If these stresses are high enough, inelastic response of the material must be considered and residual stresses will exist in the structure when the process is complete. In this paper, both thermal and transformation stresses are computed for various quenching procedures using a hollow tube for the geometric model. The relative severity of the thermal and transformation stresses and the conditions under which they occur are discussed. A general purpose finite element code, ADINA, is used for the computation.

I. INTRODUCTION. Rapid quenching of hollow steel tubes from initially high temperatures is usually undertaken to produce desired structural properties in the tubular components. This technique has been utilized in the fabrication of large caliber, long gun tubes. The initial high temperature of 843°C (1550°F) giving the steel gun tube an austentic grain structure is cooled to less than 100°C (212°F) in just a few minutes using various water spray methods. The desired end result is a martensitic structure in the material.

Due to the transient temperatures created by the water spray, the large thermal gradients in the tube wall give use to thermal stresses which change rapidly in time. In addition to these stresses, as the material grain structure transforms into martensite, a volume expansion of the structure occurs. This expansion associated with the martensite transformation gives rise to additional transient stresses, called transformation stresses. If the stresses that occur during this procedure are below the yield stress of the material, as the temperature equilibrates to room temperature and as the transformation is completed, the quenched tube is stress free, i.e., no residual stresses exist. However, if the stresses are such that some inelastic deformation occurs in the tube, then there will exist residual stresses in the tube after it is quenched. The latter is the more likely case.

This paper investigates the transient temperatures and stresses that a steel gun tube experiences in such a quenching operation. The material properties and system parameters used in the computations are chosen using the rotary forge quench facility at Watervliet Arsenal as a model. Specifically, a memorandum [1] describing the results of an experimental quench test was used to generate the approximate convection heat transfer coefficients for the study. Most of the interest is centered on the breech (or rearward) end of the tube and the muzzle (or forward) end of the tube, with emphasis on the muzzle end. This is because quench cracks are more frequent at that location. In addition to considering the two locations along the tube, the effect of altering the quench cycle is also considered. Nozzles spray water on the outer diameter of the tube as it is slowly rotated. Sprays on the breech and muzzle end can be separately controlled. The quench on the bore or inside diameter of the tube is accomplished by flushing the bore with water. It is of interest to know the effect of this flush when it occurs concurrent with the water spray on the outside diameter, or when it begins slightly before or after the OD quench.

To compute the temperatures and stresses mentioned in the above paragraph, a general purpose finite element computer program, ADINA [2], is used. The problem was treated as an axisymmetric plane strain in two dimensions. Separate analyses or runs are undertaken for the breech and muzzle sections using their respective geometries. Any tapering of the gun tube is ignored as are end effects. Although the geometry is simple, the problem is complicated due to the highly transient nature of the problem and the consideration of temperature dependent material properties and yield strength. The properties used in the report for the computations are stated in the Appendix.

II. PROBLEM STATEMENT. Thermal and transformation stresses are computed for long hollow cylinders subjected to different quench cycles. Two geometries, representing the breech and muzzle sections of a gun tube, are modeled. Effects of altering the quench cycle of the bore are considered.

III. FINITE ELEMENT PROGRAM. The finite element geometry for the problem is shown in Figure 1 along with a simplified drawing of a gun tube. Eight node quadrilateral elements were used for the model. Sixteen elements are used at the muzzle and twenty at the breech. Some preliminary work was done with other models (4 node quadrilaterals and 10 element, 8 node models) and it was decided that the current model gave sufficient accuracy and smoothness of results.

The finite element program actually consists of two parts, one for computing temperatures, ADINAT, and one for computing stresses, ADINA. Each program can stand alone, but when one wishes to compute thermal stresses using the same geometry, ADINAT produces a file which includes the temperatures at the nodes for each time step if the problem is a transient one. This ADINAT output file can then be used as input to ADINA for the stress computation.

In the program, the thermo-physical properties were considered as functions of temperature and the values are recorded in the Appendix. The convection losses during the heat transfer portion of the computation are considered to be due to the temperature difference between the tube wall and ambient, which is taken to be 18.3°C (65°F). For the computation of stresses, one has a choice of several material behavior models in ADINA. The one chosen for this work was Model 10 [3], which is applicable to the thermo-elastic-plastic solution of interest. The yield criterion assumed is the distortion energy criterion and the yield stress is assumed to be a function of temperature. No creep or hardening is assumed although the model allows both to be incorporated.

To compute thermal stresses, the problem for the transient temperatures is just solved using ADINAT as indicated above. The special file created by ADINAT is then used as input to ADINA to compute the thermal stresses. However, in many cases in solving the temperature problem, short time increments are used during periods of large transients and longer time increments when the temperature gradients are not as severe. While ADINAT allows one to change time increments, ADINA does not. This difficulty was overcome by manipulating the file used for stress computation so that with the restart capability, ADINA would see only one time increment during any one computation interval. The temperature file could also be manipulated to decrease the temperature difference between time steps with the time increment appropriately changed. Finally, the restart facility in ADINA [4] was altered so that a restart could be undertaken from any previous time instead of just the last completed step. Simple linear interpolation is used in any file manipulation.

The computation of transformation stresses and combined thermal and transformation stresses can be treated like thermal stresses with additional effort. The effect of the transformation, at least the aspect of it giving rise to stresses, is to create a volume change in the material. In this case the volume change is an increase and it occurs when the temperature at a point in the material becomes equal to the martensite start (M_s) temperature and is completed when it reaches the martensite finish (M_F) temperature. The volume change due to the martensite transformation is an expansion of about 3 to 4 percent. If the transformation is assumed to be isotropic, this represents a lineal expansion of 1 to 1-1/3 percent. The form of the term giving rise to thermal stresses is [3]

$$\alpha[T-T_{REF}]E \tag{1}$$

where: α coefficient of linear expansion
T temperature
T_{REF} a reference temperature for zero strains
E Modulus of Elasticity

Equation (1) can also be used to compute the stresses due to the transformation if α is modified.

Suppose $\bar{\alpha}$ is the modified coefficient for the transformation stress computation. Then $\bar{\alpha}$ is zero until M_s is reached. When the transformation is just complete, the temperature is M_F and a change in length per unit length, strain $\varepsilon = \Delta\ell/\ell$, is one-third the volume change or .0133. Between M_s and M_f, the transformation is assumed to progress linearly. Once M_F is reached, the expansion is assumed to be permanent. Thus, at M_s = 325°C (617°F) [5], $\bar{\alpha} = 0$ and at M_F = 260°C (500°F) [5], $\bar{\alpha}T = .0133$. Thus assuming $T_{REF} = 0$ in this case, $\bar{\alpha} = 2.666 \times 10^{-5}$. Finally, as the temperature cools below M_F, $\bar{\alpha}$ must be adjusted so that $\bar{\alpha}T$ remains constant and equal to .0133, i.e., the linear strain remains constant. If there was no inelastic deformation, when the transformation was complete, the transformation stresses should vanish as the effect of the transformation would be that of a uniform expansion. A plot of $\bar{\alpha}$ vs. temperature is shown in Figure 2.

To show the desired response, a problem was run for transformation stresses alone, with α replacing the coefficient of thermal expansion. Plastic response was suppressed by using an artifically high yield strength. The resulting stresses after the transformation was completed approached zero as they should have (due to the final uniform expansion). Because only a piecewise linear approximation to $\bar{\alpha}$ is allowed (16 points), stresses did not completely vanish, but if the 16 points are grouped immediately following the transformation, the results at that time are improved, i.e., nearer zero stress. At later times (and cooler temperatures) where the approximation was now cruder, the results worsened. This indicated that the assumptions made in computing $\bar{\alpha}$ were correct.

In a similar manner, a coefficient, α^*, can be constructed so that the combined thermal and transformation stresses could be computed. This was done by solving for α^* from

$$\alpha^*(T-T_{REF})E = \bar{\alpha}TE + \alpha(T-T_{REF})E$$

The quantities have been previously defined. Thus, the required coefficients are found from

$$\alpha^*(T) = \alpha(T) + \bar{\alpha}(T)\,\frac{T}{T-T_{REF}}$$

This is shown in Figure 2 and in the Appendix. Computations were made with each of the three coefficients. Plastic response was again suppressed. Table 1 shows some of the results. Summing the results of the separate computations for thermal stress (α) and transformation stress ($\bar{\alpha}$) should give the results for the combined solution (α^*). While this is done elastically, similar behavior should be expected in an elastic-plastic solution as computed as α^* for a new "thermal load" increment would not be affected by the type of solution.

TABLE 1. RESULTS USING COEFFICIENTS FOR COMBINED THERMAL AND TRANSFORMATION STRESS

Time (sec)	Node Location	Thermal Stress (1)	Transformation Stress (2)	Combined Stress (3)	(1) + (2)	% Diff
87.75	1,1	15,444	2,133	17,184	17,577	2.3
	7,1	-39,681	1,879	-38,165	-37,802	1.0
	16,1	85,102	1,600	86,436	86,702	.3
107.75	1,1	8,472	74,911	82,132	83,383	1.5
	7,1	-31,339	65,996	33,538	34,657	3.3
	16,1	62,566	-504,735	-442,237	-442,169	.01

NOTE: Plastic behavior is suppressed.

IV. RESULTS. Figure 3 is typical of the type of information that can result from a study of this type. It shows the response of the breech end of the tube versus temperature on a Time-Temperature-Transformation Diagram. The word typical is used above because the location of the bainite start curve is not known for the actual material modeled. The bainite start curve shown is for a steel having the composition shown in Table 2. Current gun steel composites contain alloying elements, such as vanadium, which push the bainite curve to the right. The transient cooling curves superimposed on the figure represent the cooling of the breech end of a 105 mm M68 tube. The quenching of the outer diameter was delayed 30 seconds relative to the bore quench. The convective heat transfer coefficients were found by assuming the OD temperature of the breech reached ~ 200°F in eight minutes. Assuming the bore quench to be less efficient and since no experimental temperatures were available as on the outside diameter, the convective heat transfer coefficient was taken to be one-half that of the OD. Figure 4 shows the temperature distribution throughout the wall at various times for the same geometry and quench. At the first time (2.5 seconds) one can see the bore beginning to cool and with no change on the outer diameter. The second curve shown at 32.5 seconds occurs just after the OD quench begins, and the subsequent temperature distributions show that the OD is cooled more rapidly than the bore. As the quench approaches the end of its cycle, the temperature distribution throughout the wall does not vary much. The temperature distributions vary even less in the muzzle section, e.g., Figure 8.

TABLE 2. COMPOSITION OF STEEL FOR BAINITE CURVE SHOWN

C	.31	Ni	5.07
Mn	.76	Cr	1.22
Si	.30	Mo	.48
P	.009	Al	.031
S	.023		

Figure 5 shows the variation of stress with time for a material element on the bore surface and on the outer surface. The highly transient nature and the severity of the stresses can easily be seen. Most points can be easily explained. Point 1 indicated the beginning of the quenching of the outer diameter at 30 seconds into the quench cycle. Only thermal stresses exist until point 2 is reached at about 210 seconds when the outer diameter begins to transform into martensite. At point 3, plastic deformation begins on the outer surface. At point 4, the bore begins its transformation. The bore develops plastic deformation at about 290 seconds and the bore transformation ends at point 6. The stresses beyond 350 seconds on the residual stresses that exist at those material elements after room temperature is reached. Figures 6 and 7 show the stress distribution throughout the wall for two different times. Figure 6 is included to show that the stress can vary quite strongly. The time is that when the bore first develops inelastic behavior. Figure 7 is the residual stresses that would exist after the temperature equilibrates.

The remainder of the results were generated for the muzzle section as this was an area of interest due to quench cracking under certain conditions. Four different quenches were considered:

1. No quench delay - the bore quench and OD quench are initiated simultaneously (Figures 8 through 11).

2. 30 second prebore quench - the bore quench is initiated 30 seconds before the OD quench (Figures 12 through 15).

3. 30 second postbore quench - the OD quench is initiated 30 seconds before the bore quench (Figures 16 through 19).

4. No bore quench - the bore is not quenched in this case (Figures 20 through 23).

Figures 8, 12, 16, and 20 show the temperature distribution throughout the wall for various times throughout the respective quench cycles. Except for the initial few curves, there does not appear to be great differences between the quenches. The first two or three temperature curves will indicate the type of quenching undertaken, i.e., whether the bore is quench first or not at all, etc. One can easily see however, that as time into the quench cycle progresses, the temperature distributions become rather flat. This indicates that the temperature equilibrates rapidly due to the high conductivity of the steel. This is especially true when the bore is quenched as in any of the above conditions. For the case when the bore is not quenched, there is an interesting occurrence for the parameters chosen. While the stresses on the outer surface became plastic on all the above runs, the bore saw plastic deformation only for the case of no bore quench. This will be discussed in more detail later.

The transient temperatures versus time are shown for each case for the bore area, the midsection or core, and the outer diameter in Figures 9, 13, 17, and 21. This type of curve was discussed previously. The intent of the quench is to cool the material rapidly enough to pass before the 'knee' of the bainite curve in order to form martensite structure. As mentioned previously, the bainite start curve shown is not that for the current material used, but for one with a different alloy content [5].

The following curves (Figures 10, 14, 18, 22) show the variation of stress with time for each of the quench cycles. The highly transient behavior of the problem becomes very visible. For example, in the 30 second prebore quench case (Figure 14) at 30 seconds into the quench cycle, the OD quench begins and the circumferential stress on the OD becomes tensile. At time 121 seconds, the martensitic transformation begins on the outer diameter. This material area tries to expand but it constrained by the as yet untransformed material, hence it is put into compression. At about 125 seconds, the material on the OD had yielded. The bore transformation then begins at about 133-134 seconds causing an abrupt change in stress at that point. Finally, the transformation of the OD material is completed followed by the transformation of the bore material. The material on the OD is subjected to plastic deformation (in tension) just before completion of the transformation. The bore material saw no plastic deformation at any time. Again, most of these figures are similar with the exception of Figure 22 for the case of zero bore quench. Here one can see that the stresses at the bore are larger and compressive much of the time.

The final set of figures, Figures 11, 15, 19, and 23, show the resulting residual stresses for each quench cycle. Again, these all have a similar shape with the exception of Figure 23 for the case where the bore was not quenched. As mentioned in the above paragraph, the bore developed plastic deformation only for this case. Two companies have provided steels with similar chemistry (or alloy content) for production runs of gun tubes. Tubes forged from one of the steels had a much higher frequency of quench cracking until the more severe no bore quench cycle was used. While no direct effort has been made to verify this at this time, the development of quench cracking may have been prevented by the different stress-time behavior and resulting residual stresses in this case where the bore area seems to be under compression for longer times.

V. FUTURE WORK. Concurrent efforts in the quench crack problem have been undertaken and will be providing more accurate experimental data for input to this analysis. This refers to a better estimate of the volume change in the temperature, as well as the way that the volume changes in time. An up-to-date bainite curve is being determined but preliminary indications are that it will not be in any area of the TTT diagram that would cause problems if the quenching rates are held the same. It is hoped that more accurate heat transfer coefficients can also be determined. Incorporating this new information in the analysis will certainly lead to a better understanding of the events occurring during the quenching of long hollow tubes.

REFERENCES

1. Memorandum, H. C. Sprinceam, Engineering Staff Office, Watervliet Arsenal, 1 September 1982.

2. "ADINA, Automatic Dynamic Incremental Nonlinear Analysis," Report AE81-1, September 1981, ADINA Engineering Inc., Watertown, MA.

3. M. Synder and K.-J. Bathe, "Formulation and Numerical Solution of Thermo-Elastic-Plastic and Creep Problems," NTIS: PB-274-044, June 1977.

4. Fred Gregory, Ballistics Research Laboratory, Private Communication.

5. Paul Cote, Large Caliber Weapon Systems Laboratory, Benet Weapons Laboratory, Private Communication.

6. Aerospace Structural Metals Handbook, AFML-TR-68-115.

APPENDIX

MATERIAL PROPERTIES

Table A1 shows the mechanical properties used in the program as a function of temperature. As only 16 points are allowed for describing a function, not all points were used in all calculations. The main source of the properties is noted in Reference 6. The tables shown include property values at some temperatures which were found by linear interpolation.

TABLE A1. MECHANICAL PROPERTIES AS A FUNCTION OF TEMPERATURE

Temperature (F)	Young's Modulus (x10E6 psi)	Yield Stress (psi)
65.	30.	160000.
100.	29.84	157560.
150.	29.6	154080.
200.	29.37	150060.
225.	29.26	148860.
250.	29.14	147120.
275.	29.02	145380.
300.	28.9	143640.
325.	28.78	141910.
350.	28.67	140170.
375.	28.56	138430.
400.	28.44	136680.
425.	28.32	134950.
450.	28.20	133210.
485.	28.04	130770.
500.	28.03	130750.
560.	27.67	126760.
617.	27.34	122980.
635.	27.32	122800.
1200.	27.00	88000.
1560.	18.60	65600.

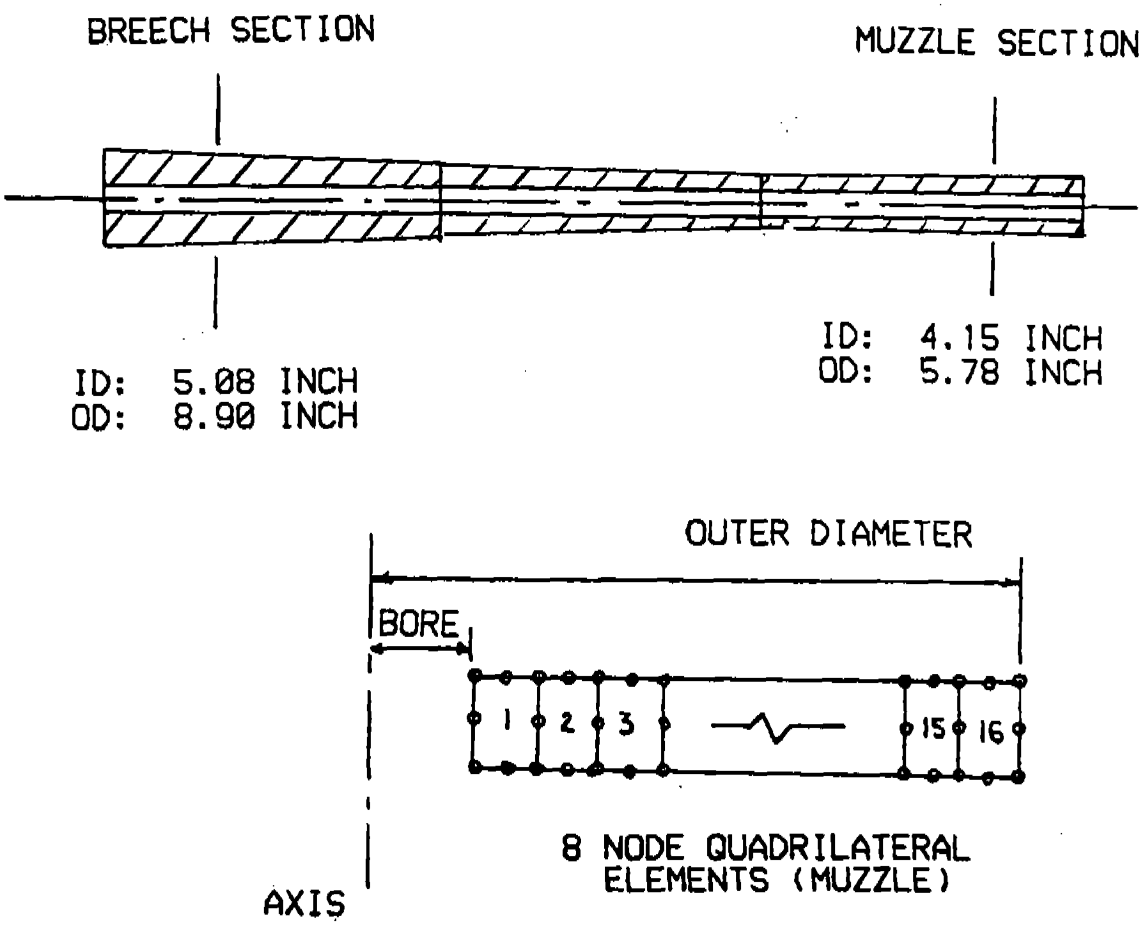
TABLE A2. THERMAL AND TRANSFORMATION COEFFICIENTS USED IN THE STRESS PROGRAM

Temperature (T)	Coefficient of Thermal Expansion (E-6 in./in.-F)	Coefficient for Transformation Computation (E-4 in./in.-F)	Coefficient for Combined Stress (E-6 in./in.-F)
65.	6.3	2.05	-2.99
100.	6.42	1.33	-3.11
150.	6.6	.89	-3.28
190.	6.75		
200.	6.78	.66	-3.47
225.	6.84	.59	-3.61
250.	6.9	.53	-3.76
275.	6.96	.48	-3.92
300.	7.02	.44	-4.08
325.	7.08	.41	-4.26
350.	7.14	.38	-4.45
375.	7.20	.36	-4.65
400.	7.27	.33	-4.86
425.	7.33	.31	-5.07
450.	7.39	.30	-5.31
485.	7.45	.27	-5.37
500.	7.51	.27	-5.82
560.	7.65	.24	-6.53
600.	7.75	.0	7.75
610.	7.76	.0	7.76
617.	7.77	.0	7.77
635.	7.8	.0	7.8
1200.	8.5	.0	8.5
1560.	6.87	.0	6.87

TABLE A3. THERMAL PROPERTIES USED AS A FUNCTION OF TEMPERATURE

Temperature (°F)	Thermal Conductivity (BTU/°F sec in.)
0	.0005
400	.000493
600	.000472
800	.000449
1000	.000412
1200	.00037
1400	.00031
1560	.00031
Temperature (°F)	**Specific Heat (BTU/#°F)**
0	.105
1300	.184
1400	.38
1550	.14

Density .284 #/in.^{3}
Martensite Start Temperature 325°C (617°F)
Martensite Finish Temperature 260°C (500°F)



PROBLEM GEOMETRY

'FIGURE 1

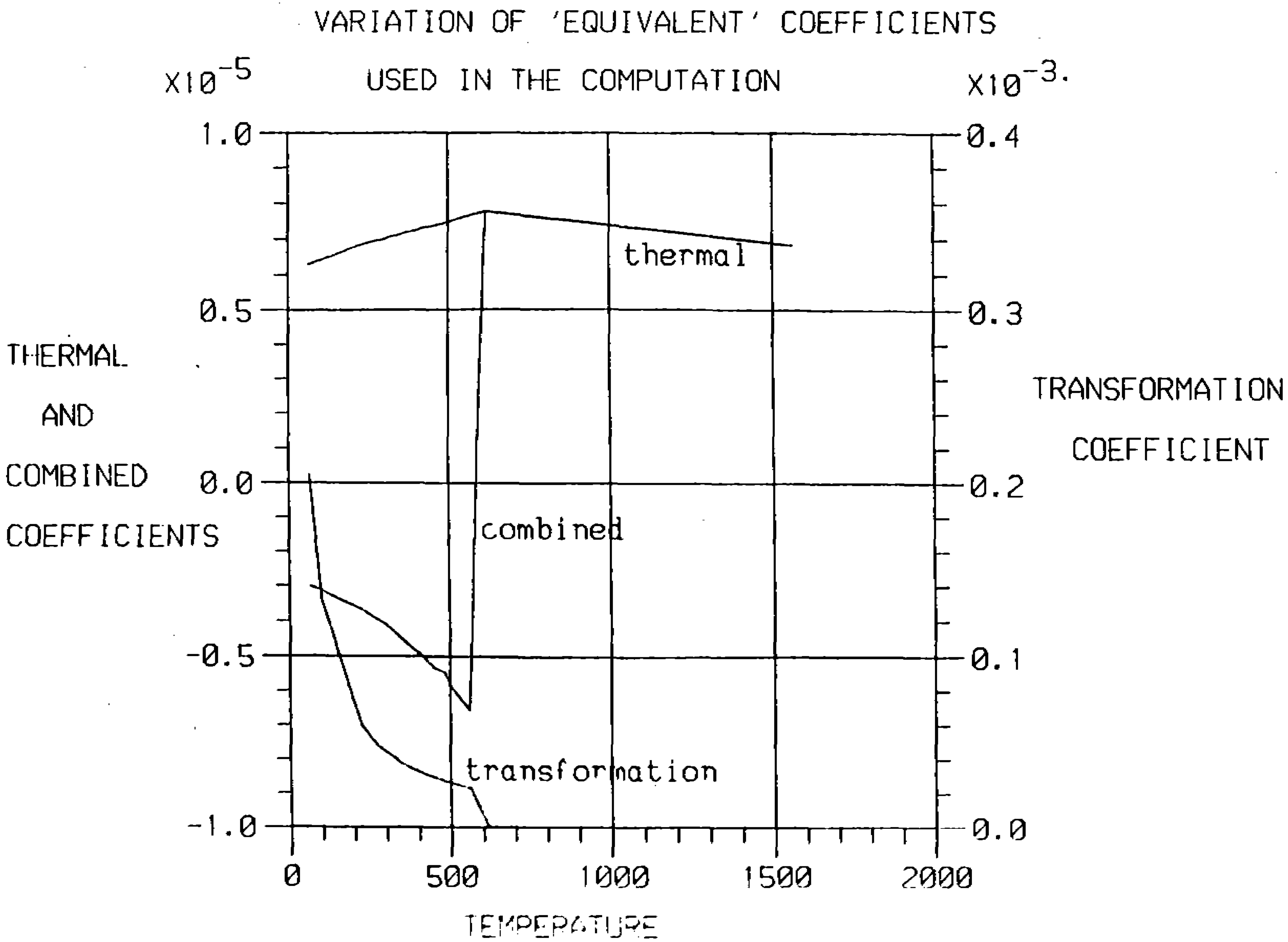

VARIATION OF 'EQUIVALENT' COEFFICIENTS
USED IN THE COMPUTATION
X10^{-5}
X10$^{-3.}$
1.0
0.5
0.0
-0.5
-1.0
0.4
0.3
0.2
0.1
0.0
THERMAL AND COMBINED COEFFICIENTS
TRANSFORMATION COEFFICIENT
thermal
combined
transformation
0
500
1000
1500
2000
TEMPERATURE

FIGURE 2

TTT DIAGRAM - TRANSIENT COOLING FOR BREECH END
30 SECOND PREBORE QUENCH

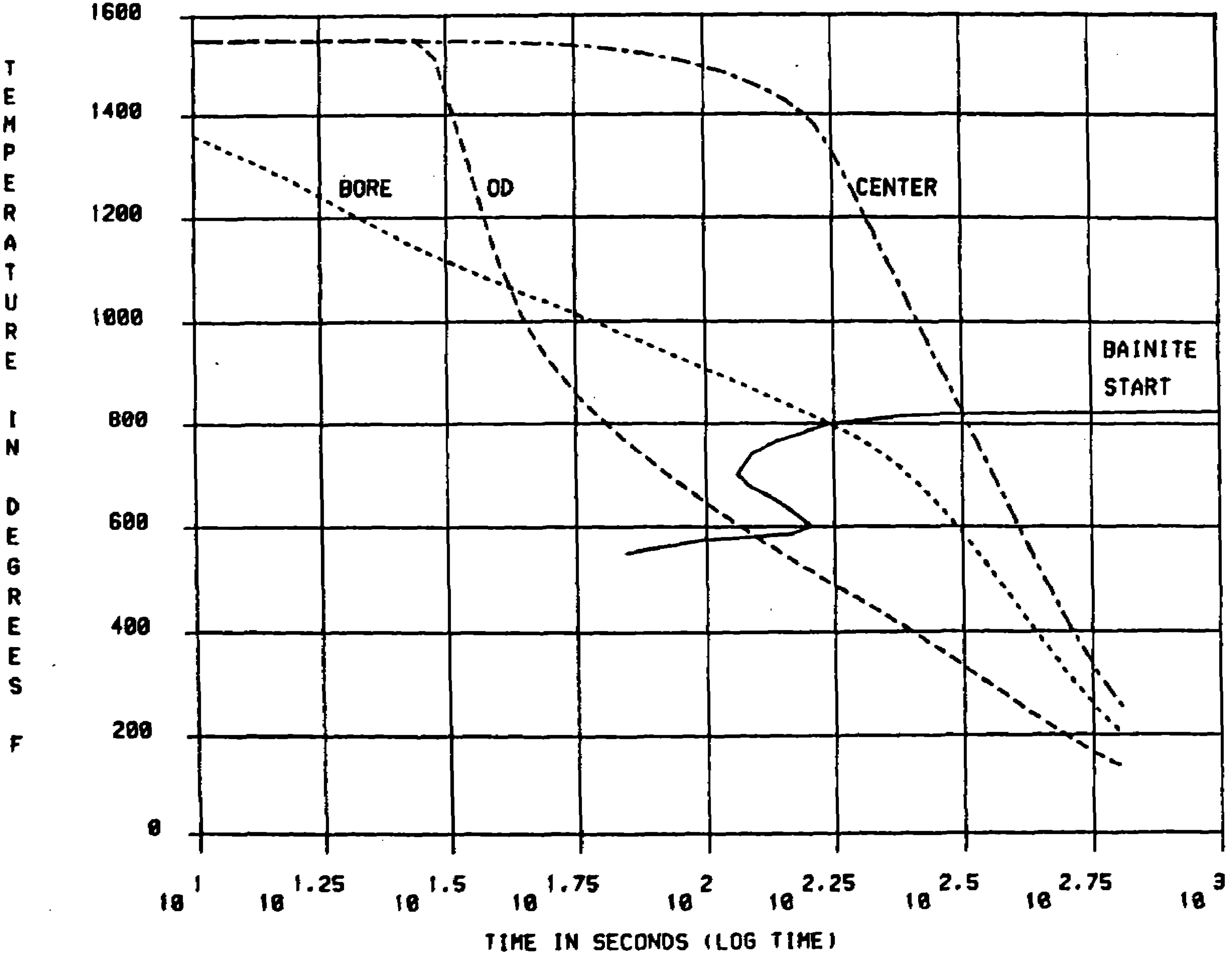


FIGURE 3

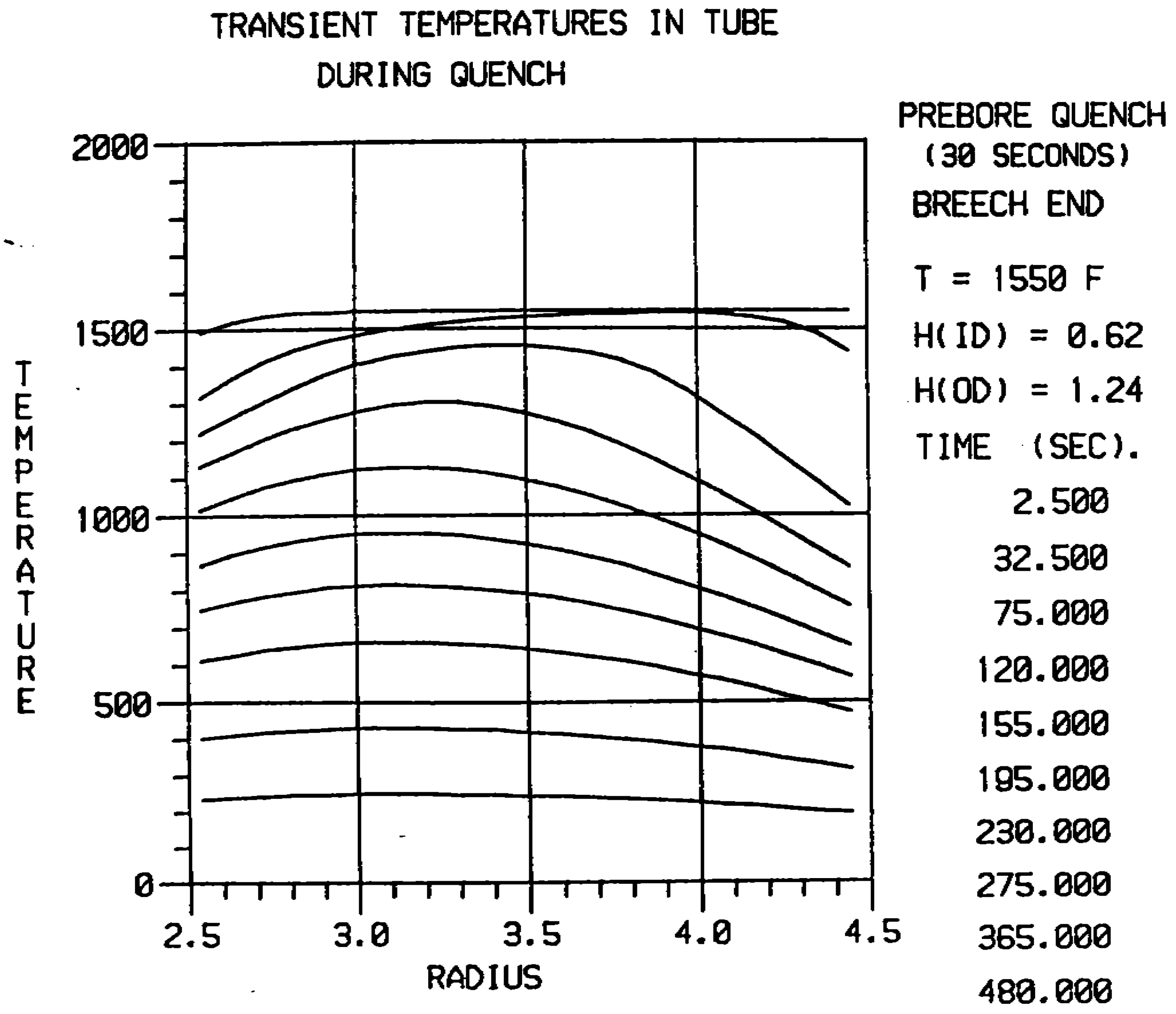

TRANSIENT TEMPERATURES IN TUBE DURING QUENCH

FIGURE 4

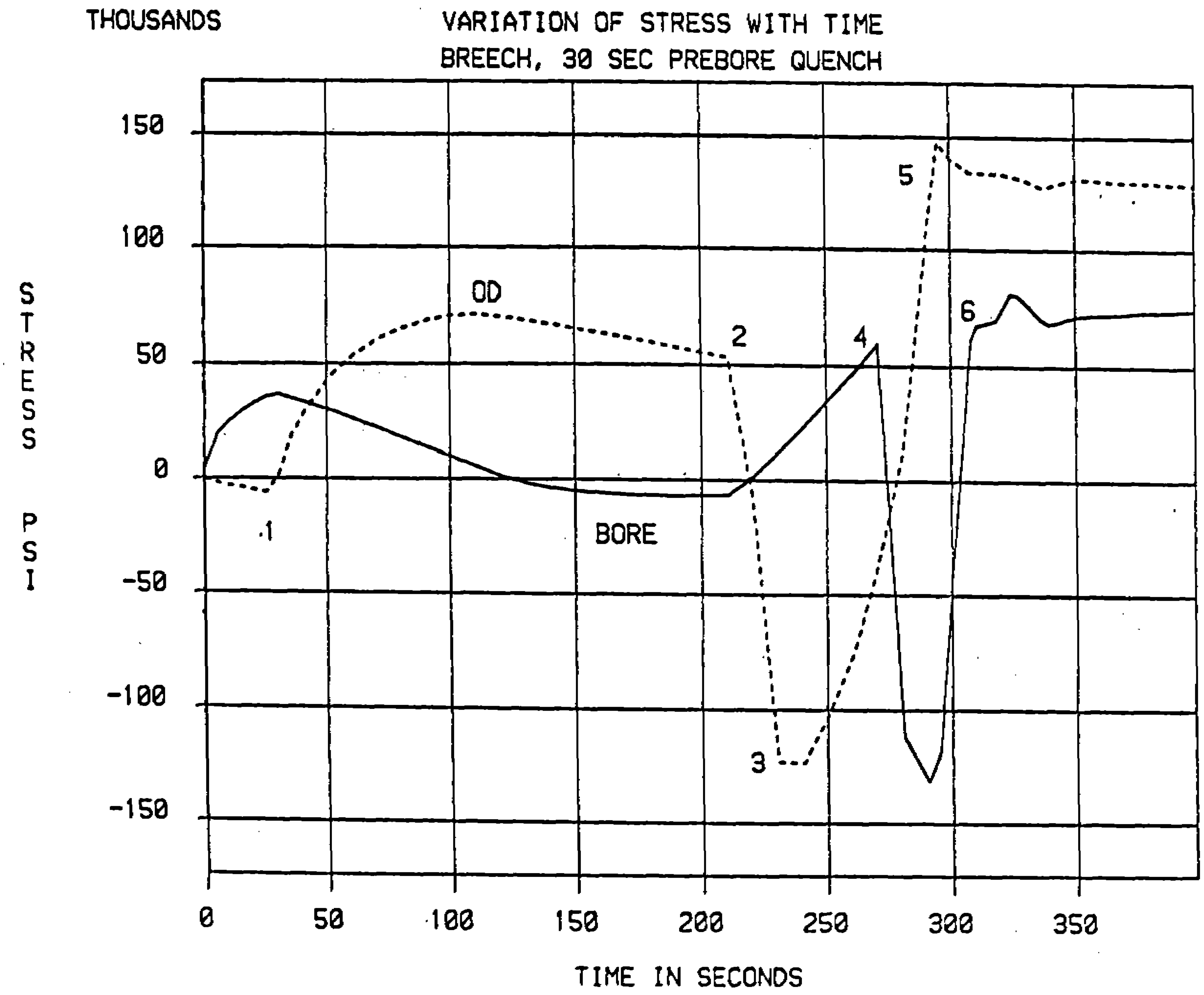

THOUSANDS
VARIATION OF STRESS WITH TIME
BREECH, 30 SEC PREBORE QUENCH
150
100
50
0
-50
-100
-150
STRESS
PSI
OD
BORE
1
2
3
4
5
6
0
50
100
150
200
250
300
350
TIME IN SECONDS

FIGURE 5

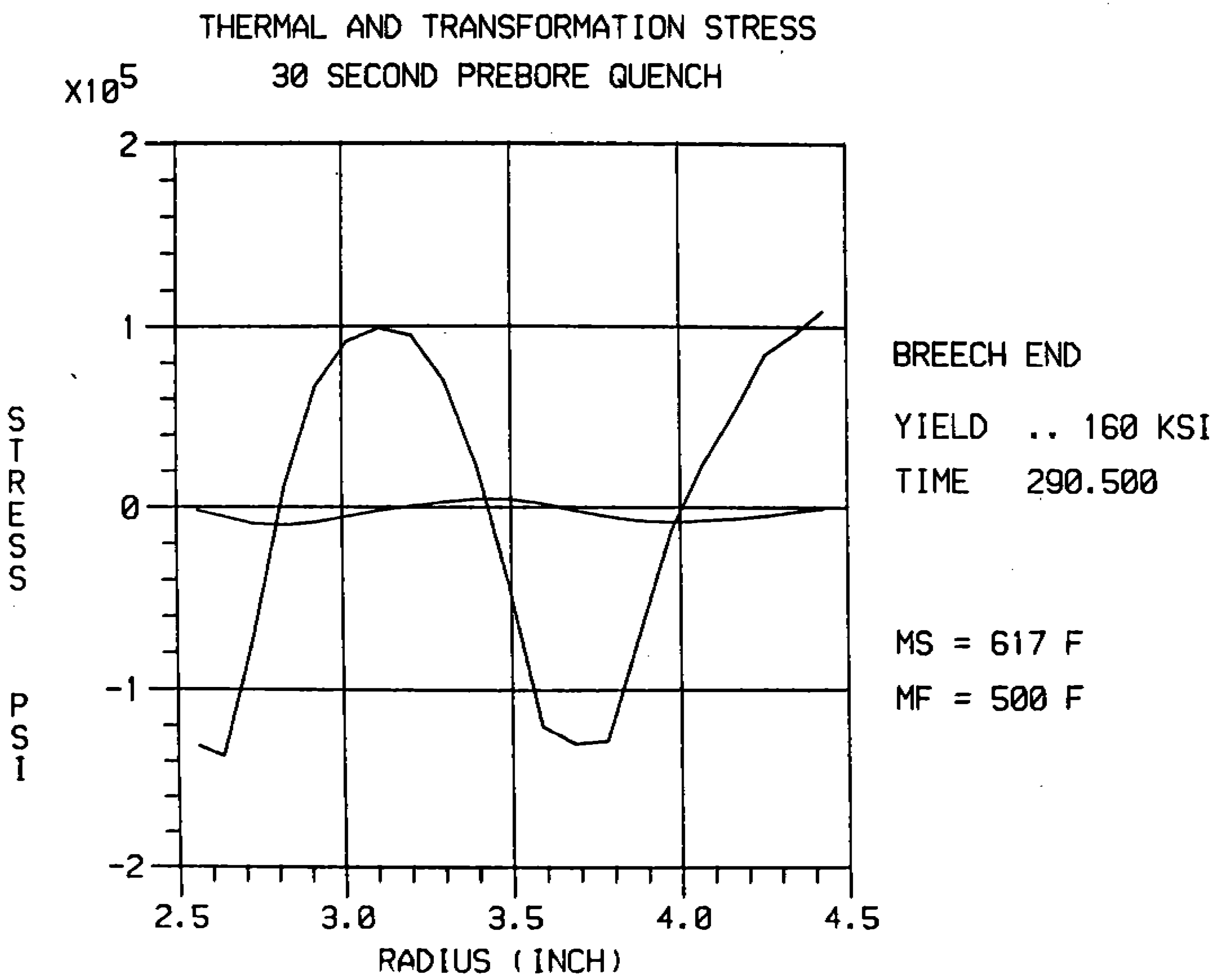
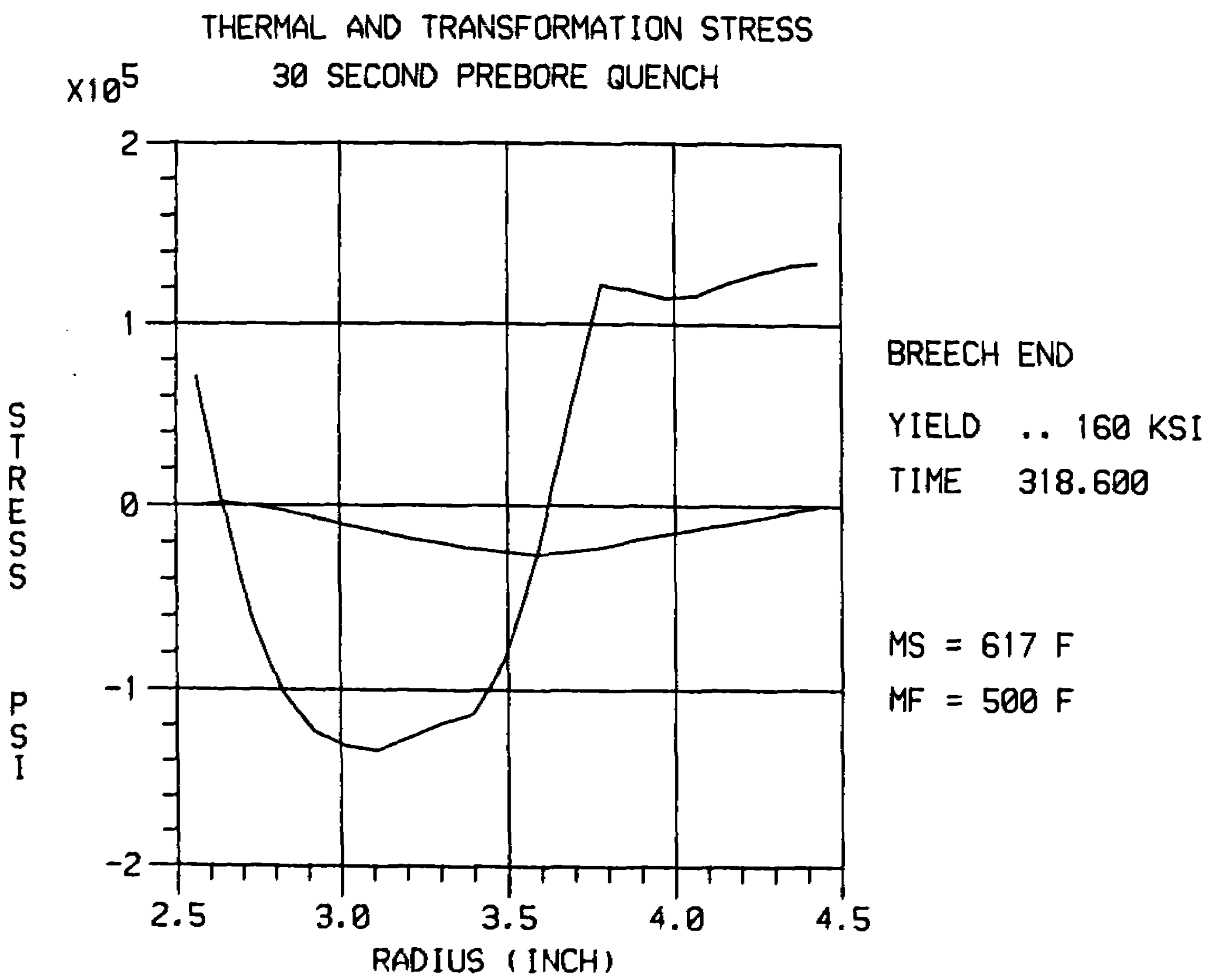

THERMAL AND TRANSFORMATION STRESS
30 SECOND PREBORE QUENCH
X10^5
2
1
0
-1
-2
STRESS
PSI
2.5
3.0
3.5
4.0
4.5
RADIUS (INCH)
BREECH END
YIELD .. 160 KSI
TIME 290.500
MS = 617 F
MF = 500 F

FIGURE 6



THERMAL AND TRANSFORMATION STRESS
30 SECOND PREBORE QUENCH

FIGURE 7

TRANSIENT TEMPERATURES IN TUBE
DURING QUENCH

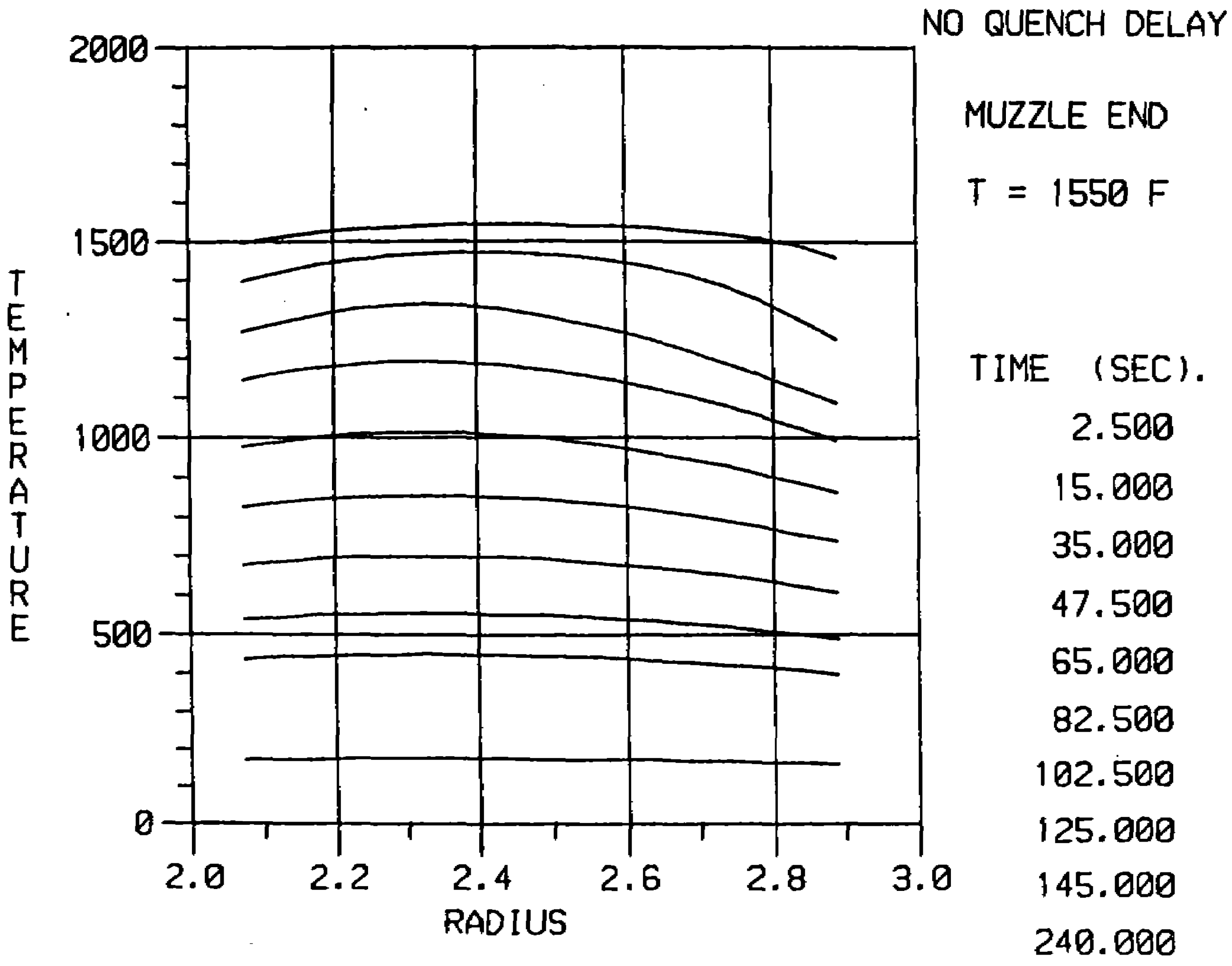


FIGURE 8

TTT DIAGRAM - TRANSIENT COOLING FOR MUZZLE END
(NO QUENCH DELAY)

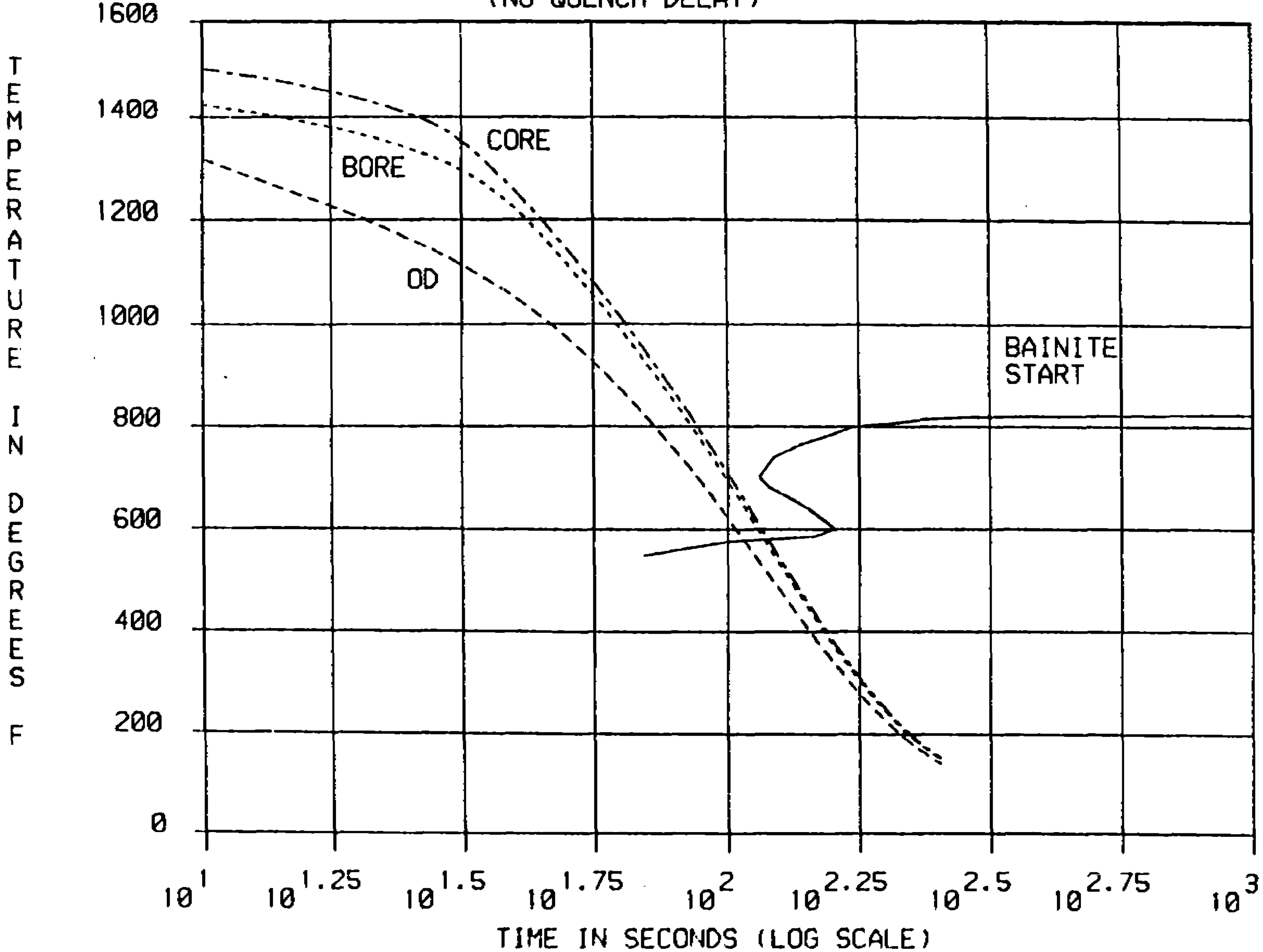


FIGURE 9

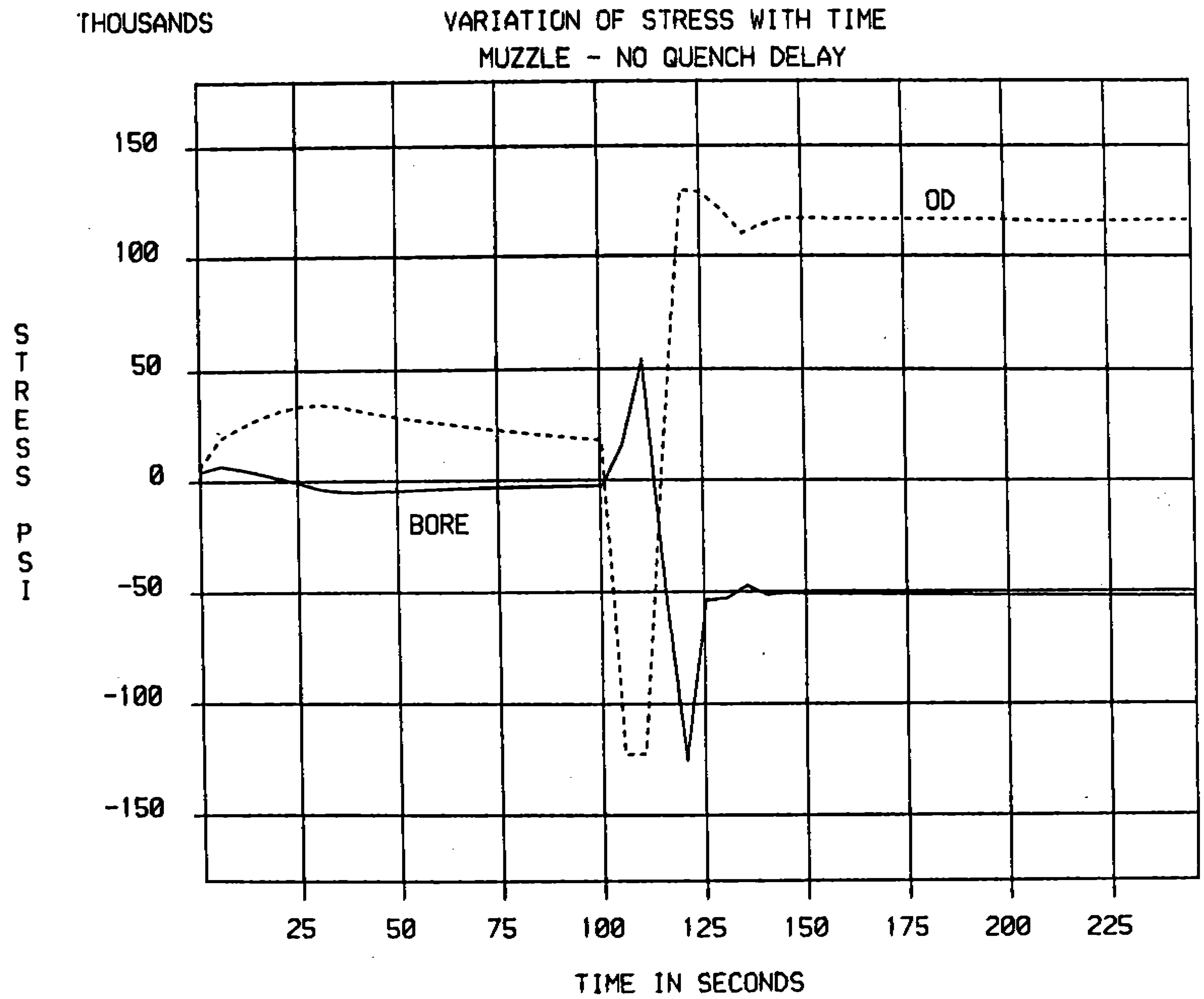

THOUSANDS
VARIATION OF STRESS WITH TIME
MUZZLE - NO QUENCH DELAY
150
100
50
0
-50
-100
-150
STRESS PSI
OD
BORE
25
50
75
100
125
150
175
200
225
TIME IN SECONDS

FIGURE 10

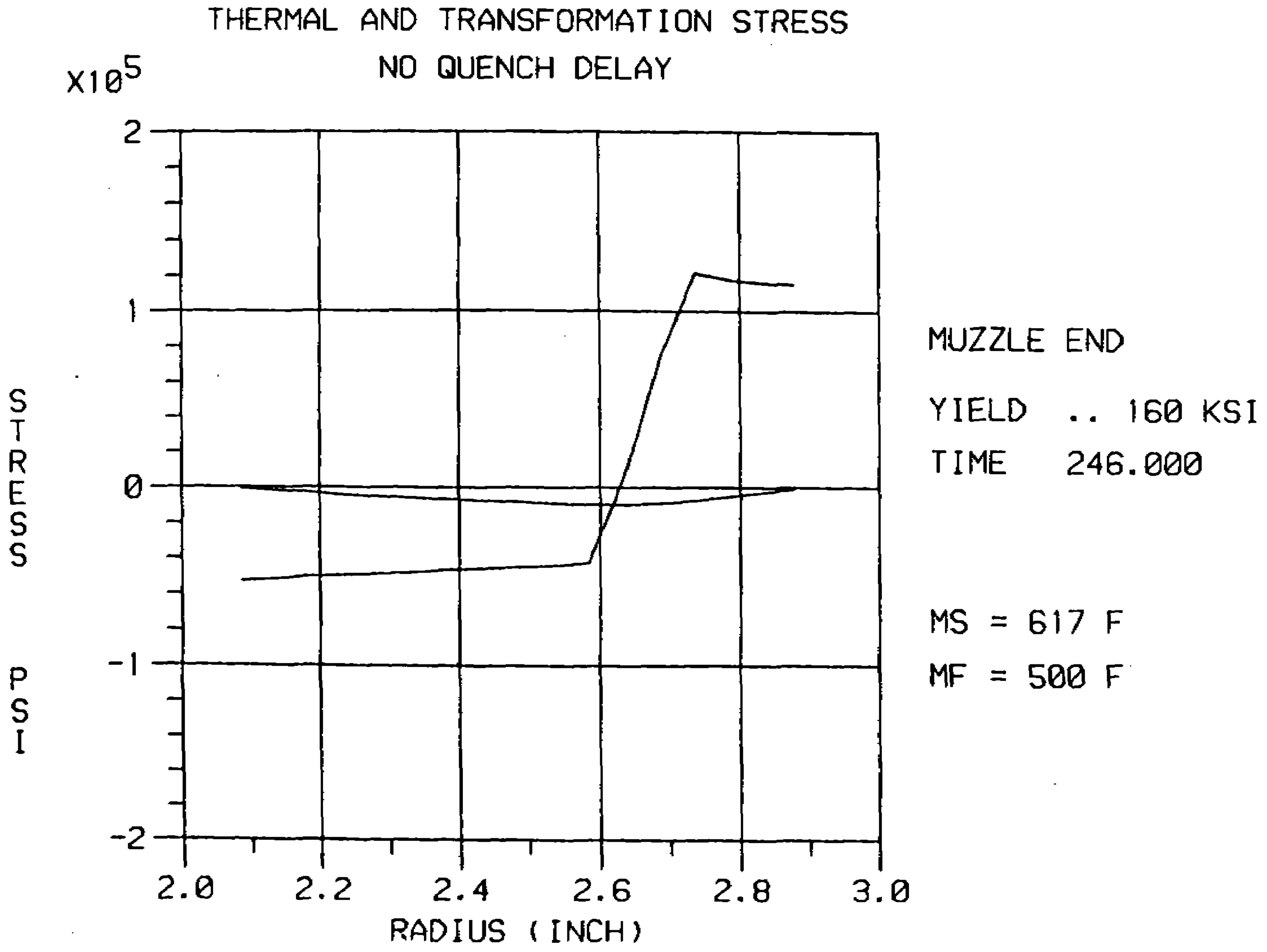

THERMAL AND TRANSFORMATION STRESS
NO QUENCH DELAY
X10^5
2
1
0
-1
-2
STRESS PSI
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS (INCH)
MUZZLE END
YIELD .. 160 KSI
TIME 246.000
MS = 617 F
MF = 500 F

FIGURE 11

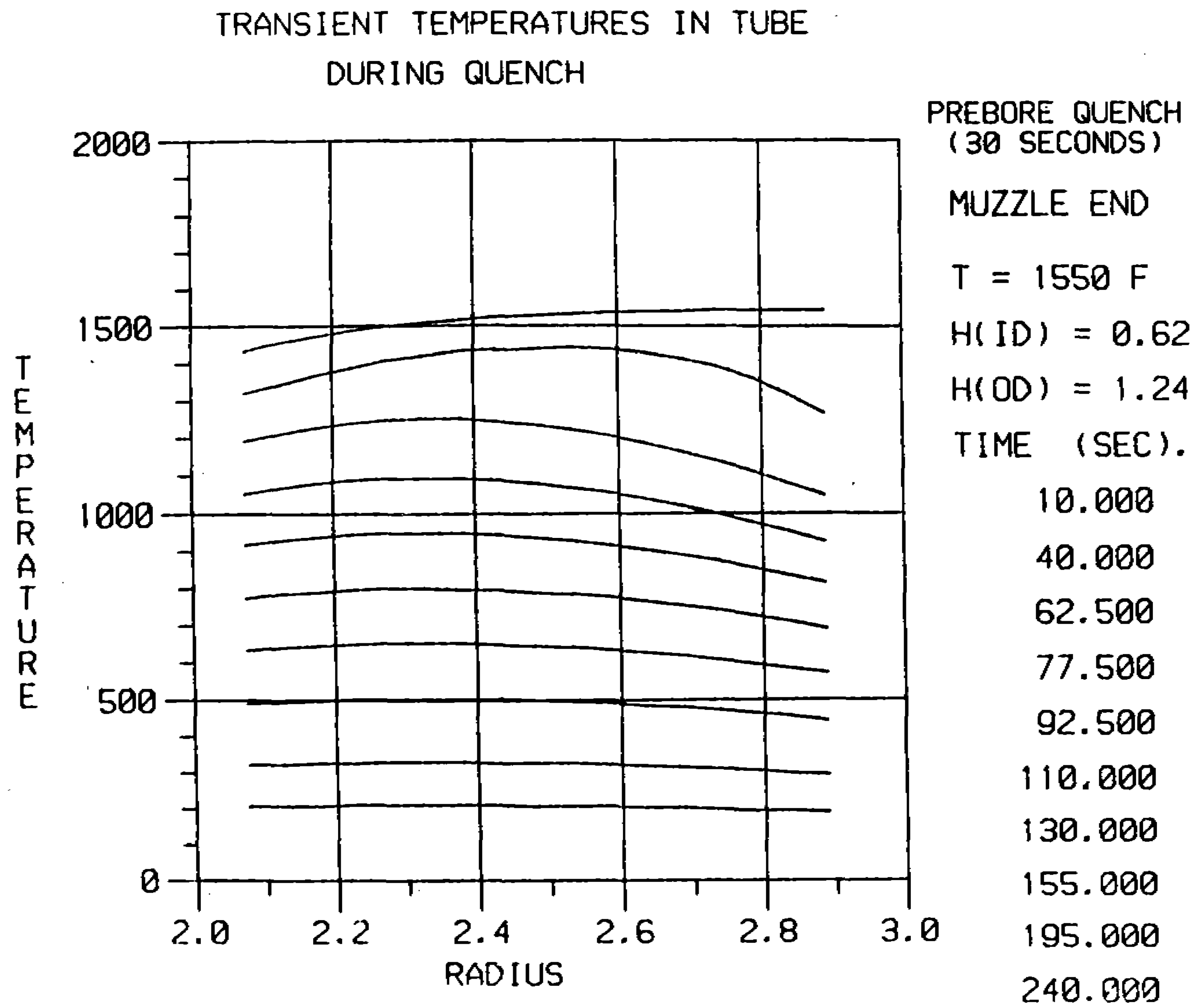

TRANSIENT TEMPERATURES IN TUBE
DURING QUENCH
PREBORE QUENCH
(30 SECONDS)
MUZZLE END
T = 1550 F
H(ID) = 0.62
H(OD) = 1.24
TIME (SEC).
10.000
40.000
62.500
77.500
92.500
110.000
130.000
155.000
195.000
240.000
TEMPERATURE
2000
1500
1000
500
0
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS

FIGURE 12

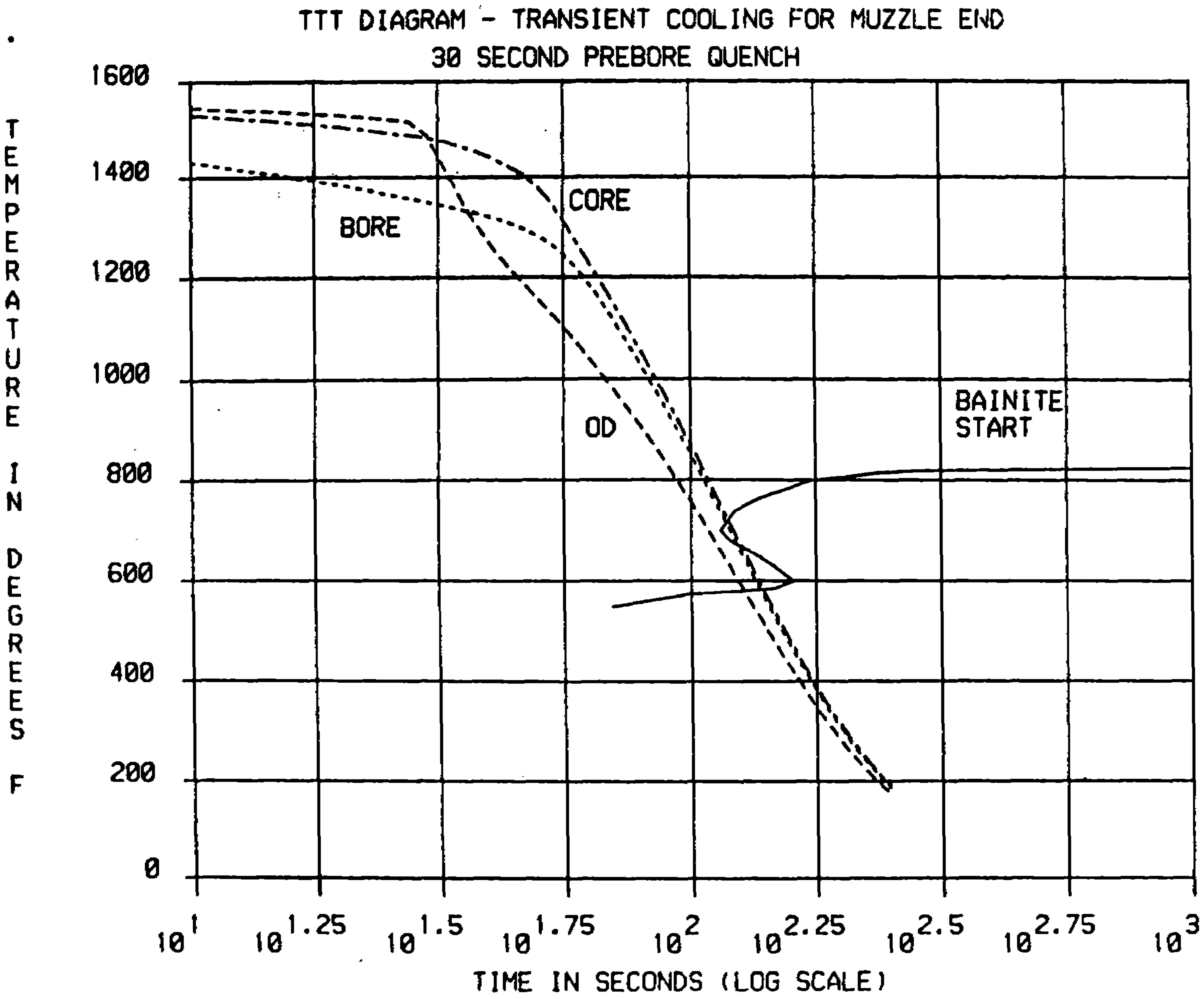

TTT DIAGRAM - TRANSIENT COOLING FOR MUZZLE END
30 SECOND PREBORE QUENCH
TEMPERATURE IN DEGREES F
1600
1400
1200
1000
800
600
400
200
0
CORE
BORE
OD
BAINITE START
10^1
$10^{1.25}$
$10^{1.5}$
$10^{1.75}$
10^2
$10^{2.25}$
$10^{2.5}$
$10^{2.75}$
10^3
TIME IN SECONDS (LOG SCALE)

FIGURE 13

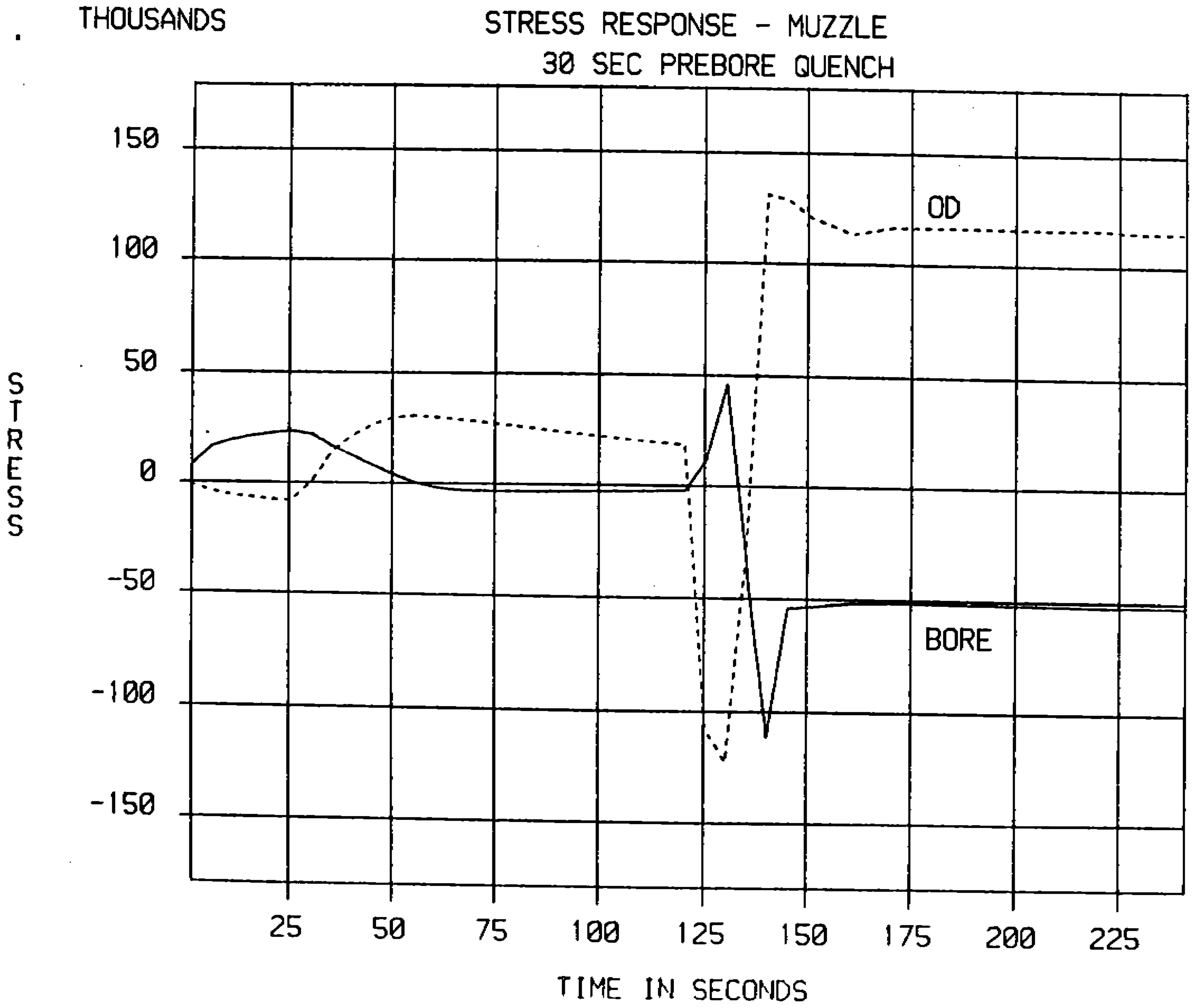
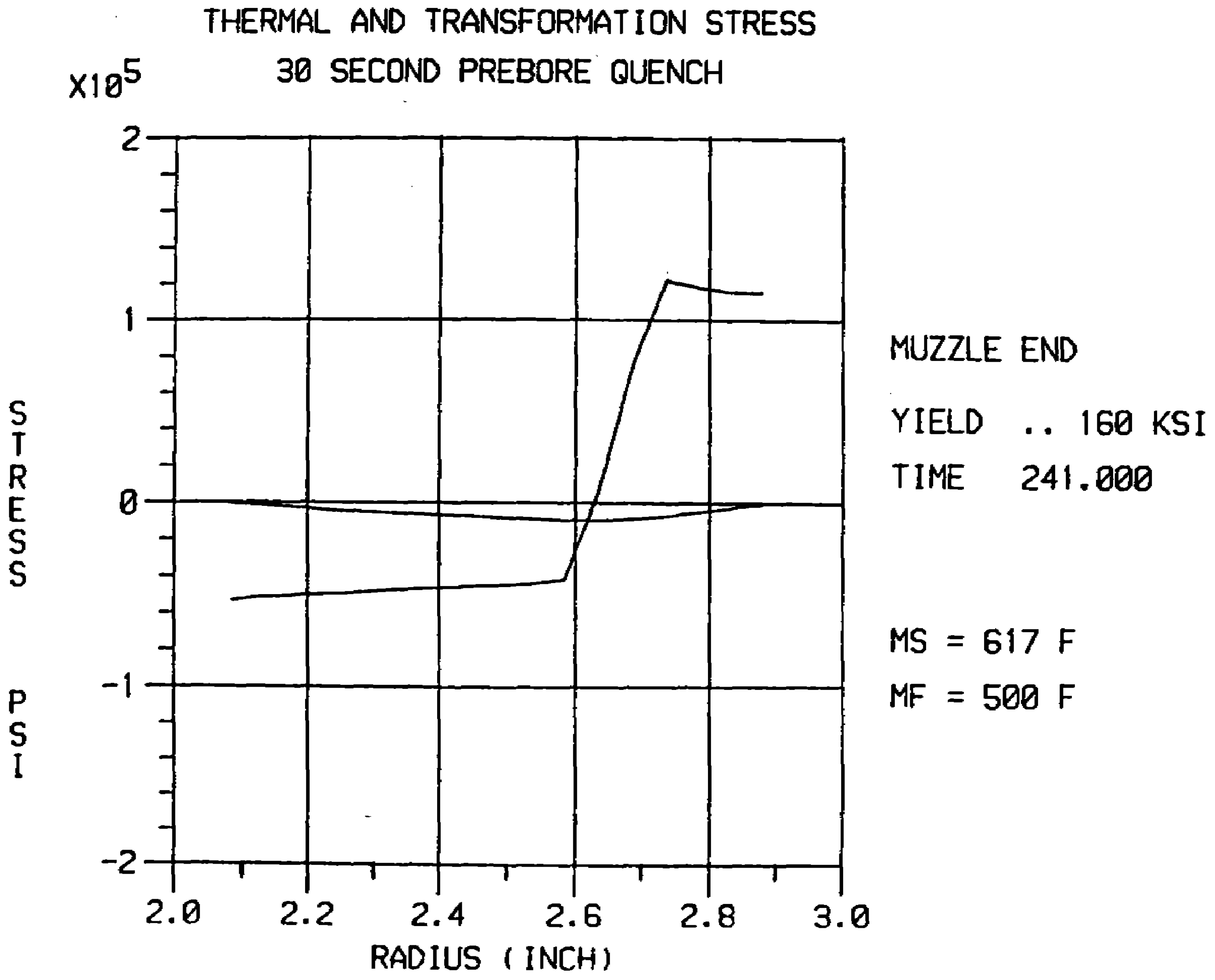

THOUSANDS
STRESS RESPONSE - MUZZLE
30 SEC PREBORE QUENCH
STRESS
150
100
50
0
-50
-100
-150
OD
BORE
25
50
75
100
125
150
175
200
225
TIME IN SECONDS

FIGURE 14



THERMAL AND TRANSFORMATION STRESS
30 SECOND PREBORE QUENCH

FIGURE 15

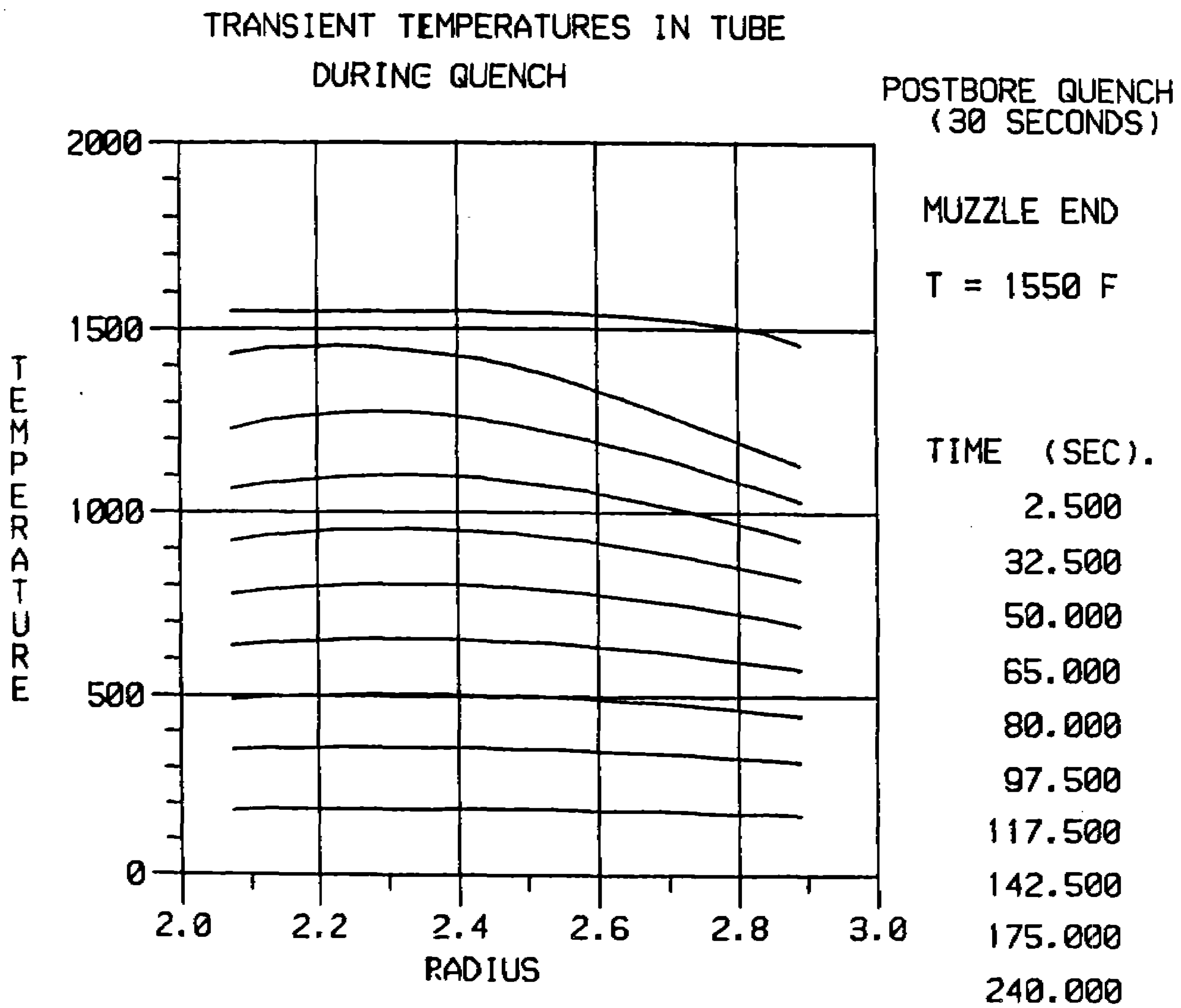

TRANSIENT TEMPERATURES IN TUBE
DURING QUENCH
POSTBORE QUENCH
(30 SECONDS)
MUZZLE END
T = 1550 F
TIME (SEC).
2.500
32.500
50.000
65.000
80.000
97.500
117.500
142.500
175.000
240.000
TEMPERATURE
2000
1500
1000
500
0
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS

FIGURE 16

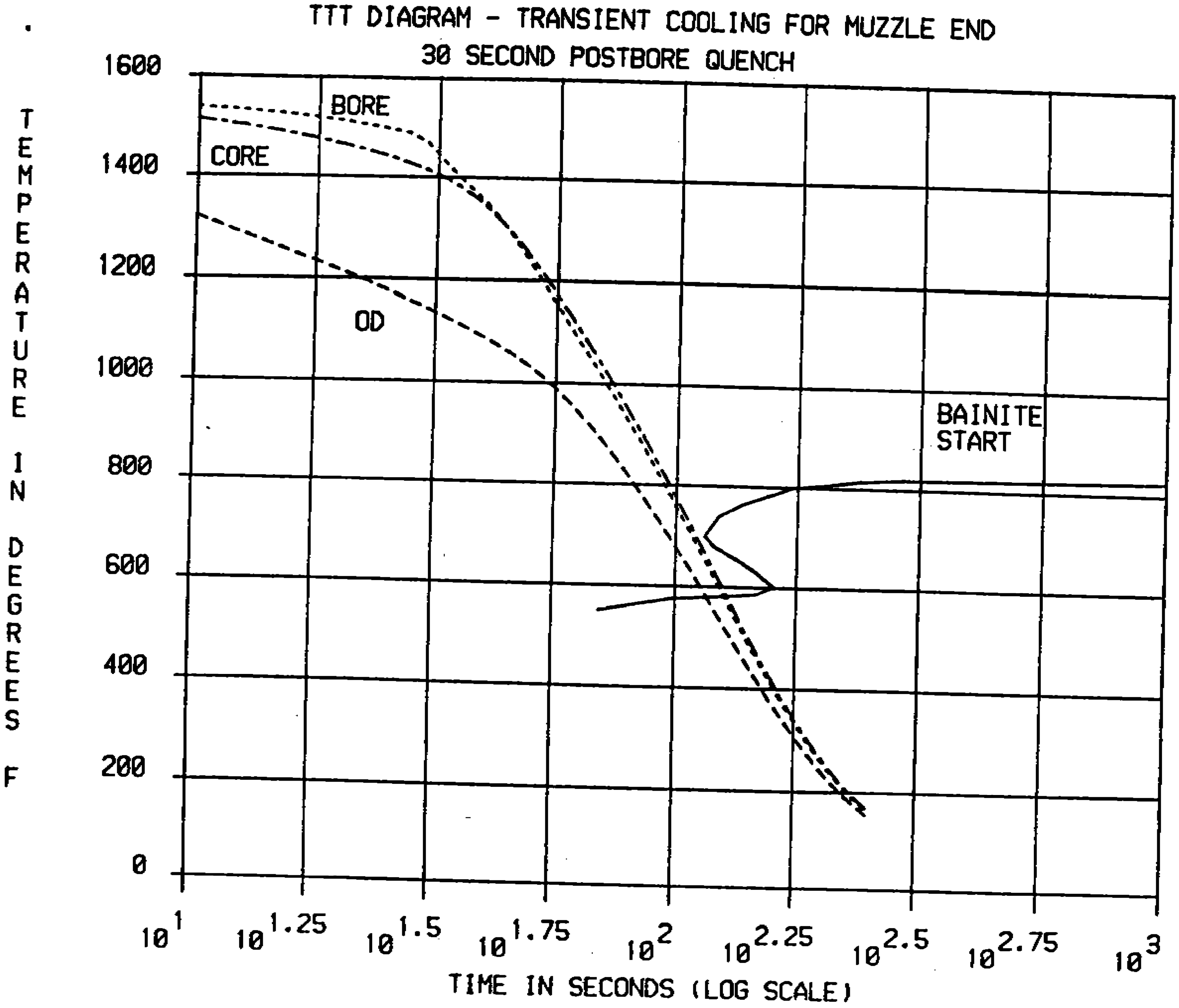

TTT DIAGRAM - TRANSIENT COOLING FOR MUZZLE END
30 SECOND POSTBORE QUENCH
TEMPERATURE IN DEGREES F
1600
1400
1200
1000
800
600
400
200
0
BORE
CORE
OD
BAINITE START
10^1
$10^{1.25}$
$10^{1.5}$
$10^{1.75}$
10^2
$10^{2.25}$
$10^{2.5}$
$10^{2.75}$
10^3
TIME IN SECONDS (LOG SCALE)

FIGURE 17

VARIATION OF STRESS WITH TIME
30 SECOND POSTBORE QUENCH

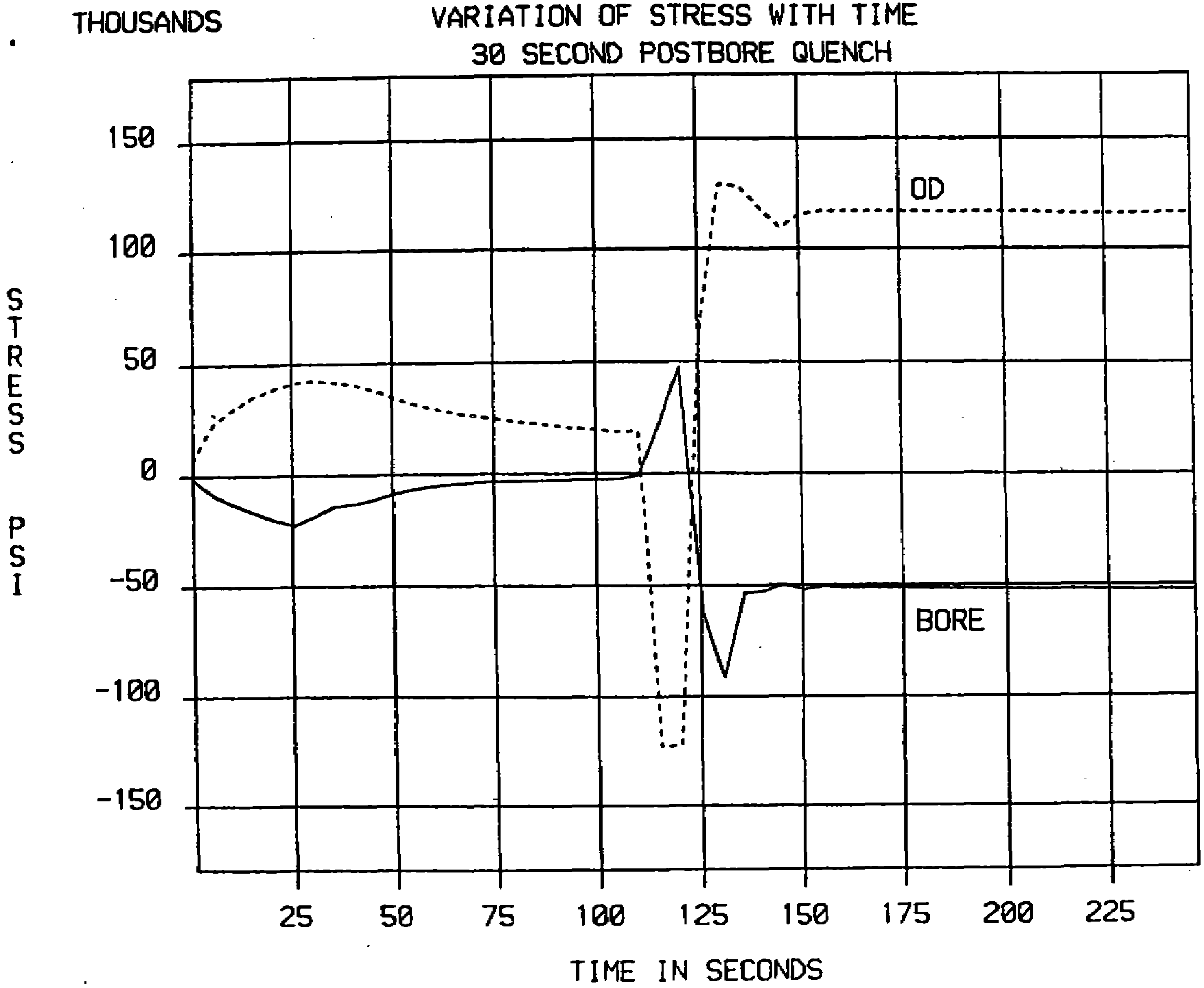


FIGURE 18

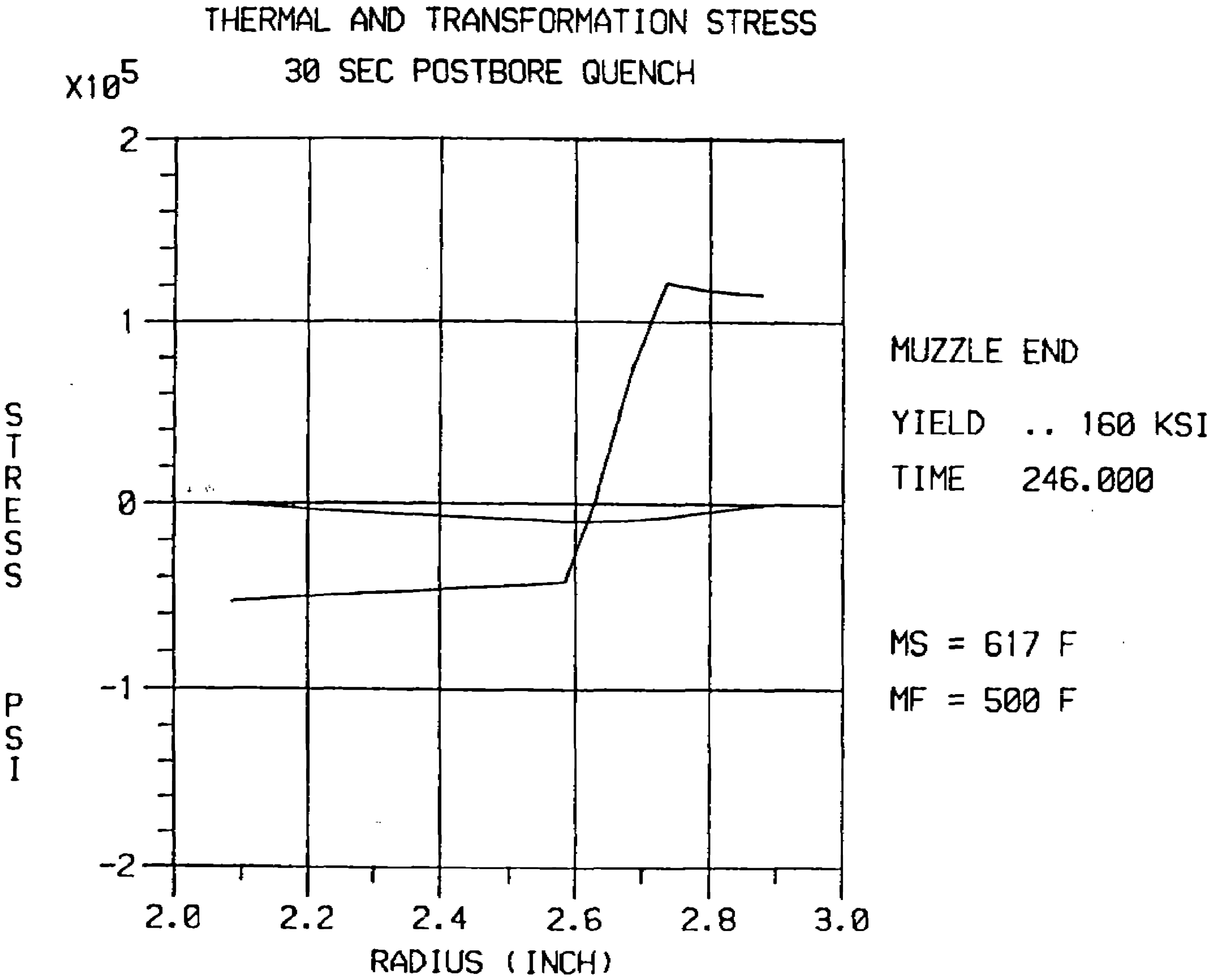

THERMAL AND TRANSFORMATION STRESS
30 SEC POSTBORE QUENCH
X10^5
2
1
0
-1
-2
STRESS
PSI
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS (INCH)
MUZZLE END
YIELD .. 160 KSI
TIME 246.000
MS = 617 F
MF = 500 F

FIGURE 19

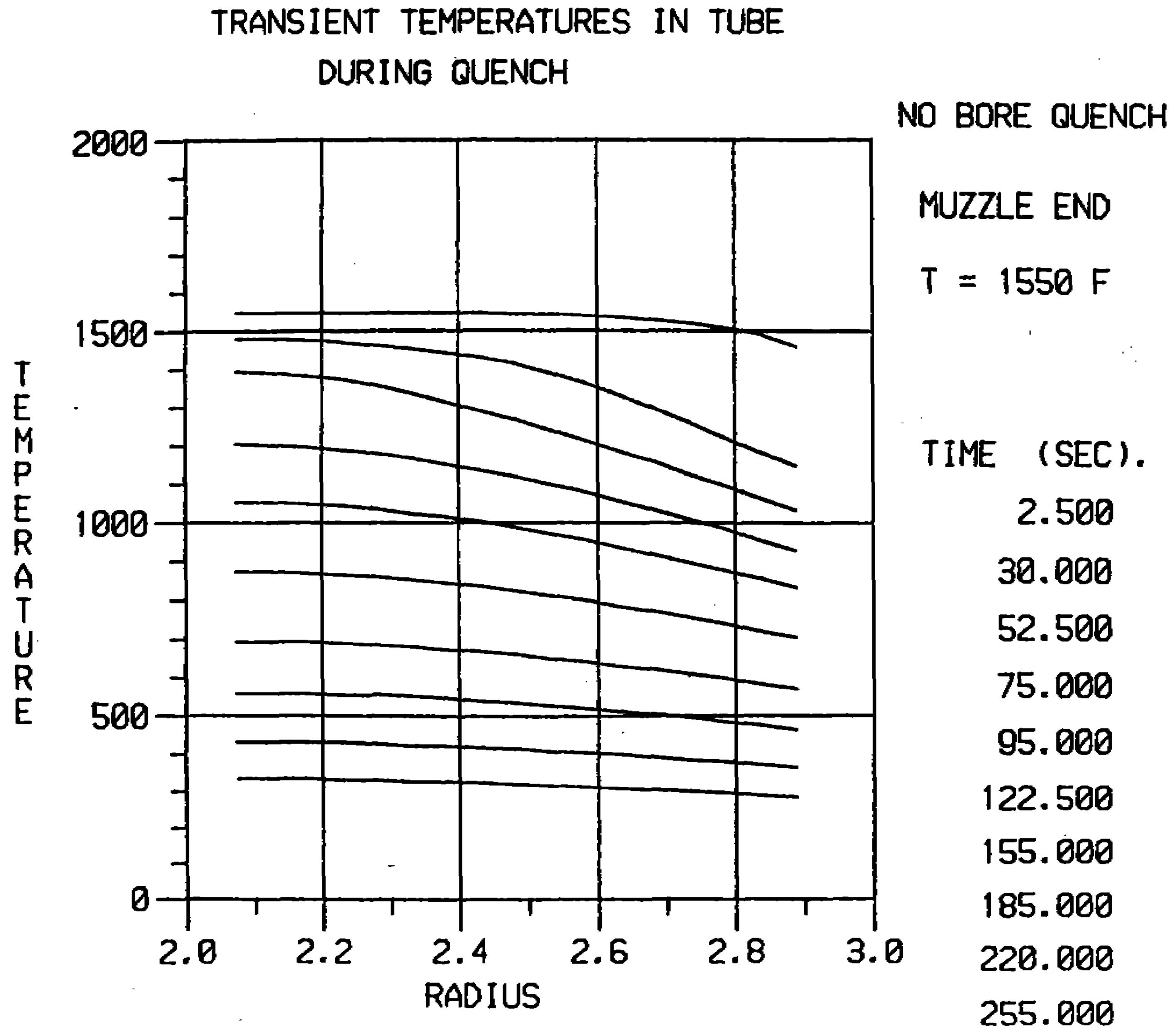

TRANSIENT TEMPERATURES IN TUBE
DURING QUENCH
NO BORE QUENCH
MUZZLE END
T = 1550 F
TEMPERATURE
2000
1500
1000
500
0
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS
TIME (SEC).
2.500
30.000
52.500
75.000
95.000
122.500
155.000
185.000
220.000
255.000

FIGURE 20

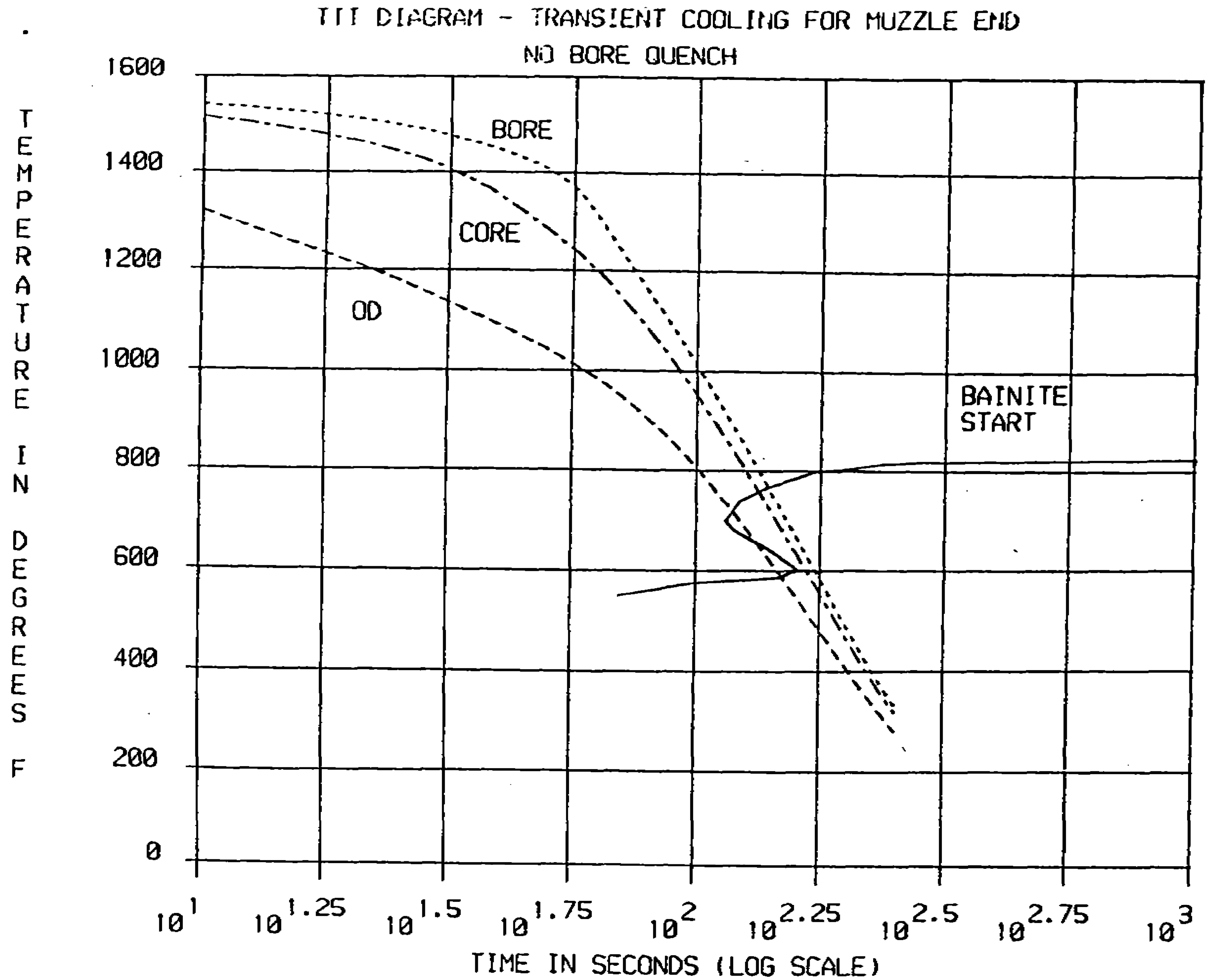

TTT DIAGRAM - TRANSIENT COOLING FOR MUZZLE END
NO BORE QUENCH
TEMPERATURE IN DEGREES F
1600
1400
1200
1000
800
600
400
200
0
BORE
CORE
OD
BAINITE START
10^{1}
$10^{1.25}$
$10^{1.5}$
$10^{1.75}$
10^{2}
$10^{2.25}$
$10^{2.5}$
$10^{2.75}$
10^{3}
TIME IN SECONDS (LOG SCALE)

FIGURE 21

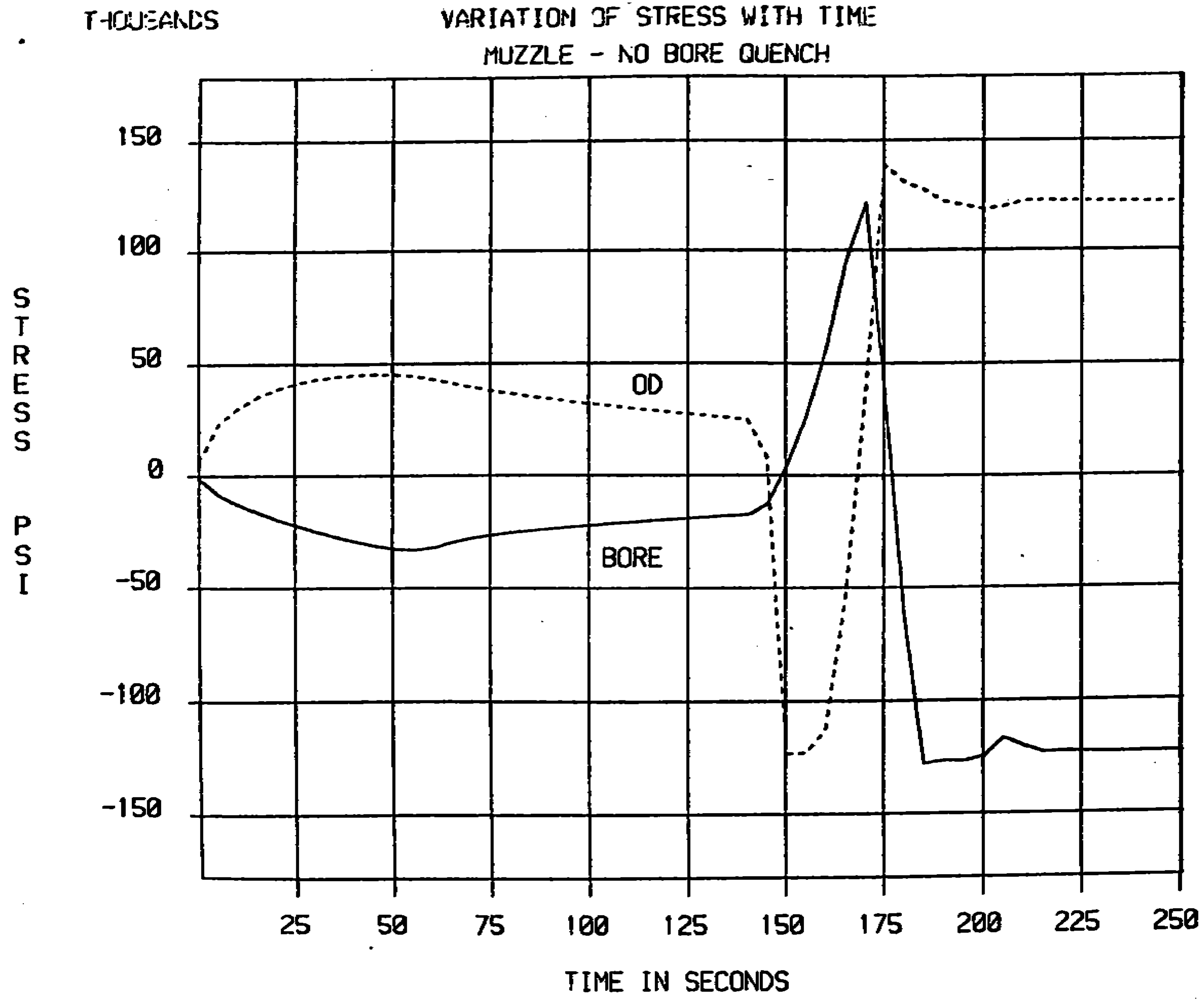

THOUSANDS
VARIATION OF STRESS WITH TIME
MUZZLE - NO BORE QUENCH
STRESS PSI
150
100
50
0
-50
-100
-150
OD
BORE
25
50
75
100
125
150
175
200
225
250
TIME IN SECONDS

FIGURE 22

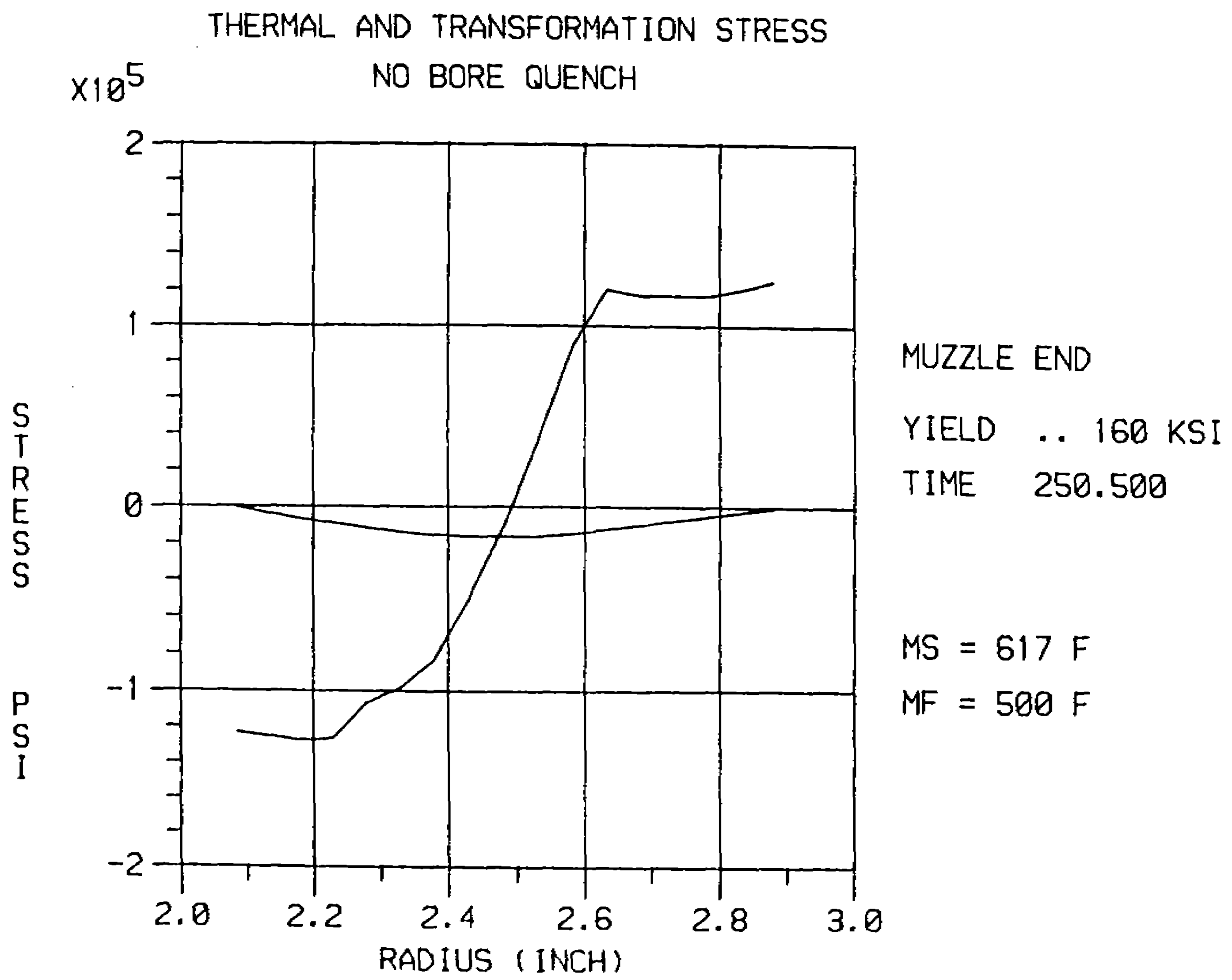

THERMAL AND TRANSFORMATION STRESS
NO BORE QUENCH
X10^5
STRESS
PSI
2
1
0
-1
-2
2.0
2.2
2.4
2.6
2.8
3.0
RADIUS (INCH)
MUZZLE END
YIELD .. 160 KSI
TIME 250.500
MS = 617 F
MF = 500 F

FIGURE 23

FINITE ELEMENT ANALYSIS OF FABRICS WITH NONLINEAR STRESS-STRAIN LAWS

A. R. Johnson
US Army Materials and Mechanics Research Center (DRXMR-SMM)
Watertown, MA 02172

ABSTRACT. The material constitutive relationships for woven fabrics are nonlinear. This, together with the fact that transversely loaded fabric membranes undergo large deformations, presents a difficult analysis problem to designers using fabrics in structures. In this effort, the constant strain triangular element is used for the development of finite element gradient and tangent matrices for the case when the material energy density functional is a C^2 function of the warp and fill strains. The element matrices are derived using the nonlinear Green strain-displacement relations to describe the warp and fill strains in terms of the deformations of the fabric. Biaxial stress-strain data for a 2.6 oz/yd^2 cotton cloth is used to obtain approximate stress-strain functionals and an energy density functional. The element is used to determine the deformations and stresses in a uniformly loaded square fabric membrane.

INTRODUCTION. Fabrics are used as structural components in lighter than air vehicles, parachutes, and tents. Recent works[1-5] have concentrated on determining material properties, constitutive relations and constructing appropiate finite element algorithms to deal with the large deformations and nonlinear materials. Previous constitutive relationships used stresses as independent variables along with warp (lengthwise)-to-fill (woof) stress ratio.[1-3] This presents a complication to displacement finite element formulations.

In this effort the idea of fitting a function to the stress-strain data with the strains as independent variables is pursued with a finite element formulation similar to that used by Johnson[6,7] to determine the deformations of rubber membranes. The warp and fill stresses are expressed via warp and fill strains. An energy density functional for fabric membranes is then obtained in terms of the fabric warp and fill strains. Green's strain-displacement relations are used and a potential energy functional is obtained in terms of material displacements. A total Lagrangian formulation is then constructed and used to obtain displacements, strains and stresses in a uniformly loaded initially flat square fabric membrane.

CONSTITUTIVE RELATION. The stress-strain data for biaxially loaded woven fabrics indicate that the response of a fabric, say in the warp direction, is dependent on the ratio of the warp-to-fill stresses (see Refs 2, 3). In Figure 1 some of the biaxial stress-strain data from reference 2 is presented for a 2.6 oz/yd^2 cotton cloth. This data was used to generate the curves shown in Figure 2. When similar curves are drawn for linear materials, a series of equally spaced constant stress contours is obtained whose slopes in the ϵ_1, ϵ_2 plane are dependent on Poisson's ratio. Comparisons indicate that the fabric constant load profiles are associated with a variable Poisson's ratio, and that the magnitude of the stresses grows ever more rapidly for fabrics as the strains increase. That is, for fabrics, the spacing between the constant load profiles for a given load change tends to decrease as the load increases, while it remains constant for linear materials. Then, a constitutive relationship for fabrics should have a variable Poisson's ratio and an increasing load-to-strain slope as the strain increases. After investigating several possible functions, an exponential form was selected, i.e.,

$$N_1 = e^{C_{11}\left[\epsilon_1 + V_1(\epsilon_1,\epsilon_2)\epsilon_2\right]} - 1$$
$$N_2 = e^{C_{22}\left[\epsilon_2 + V_2(\epsilon_1,\epsilon_2)\epsilon_1\right]} - 1 \qquad (1)$$

Upon expanding the exponentials in (1) we obtain

$$N_1 = C_{11}\epsilon_1 + V_1 C_{11}\epsilon_2 + \frac{C_{11}^2}{2}\left[\epsilon_1 + V_1\epsilon_2\right]^2 + \cdots$$
$$N_2 = C_{22}V_2\epsilon_1 + C_{22}\epsilon_2 + \frac{C_{22}^2}{2}\left[V_2\epsilon_1 + \epsilon_2\right]^2 + \cdots \qquad (2)$$

The first order terms in (2), with V_1, V_2 constant, represent a linear material. With suitable constants C_{11}, C_{22} and functions $V_1(\epsilon_1,\epsilon_2)$, $V_2(\epsilon_1,\epsilon_2)$ we may expect to obtain a behavior close to that of a fabric.

A material which strictly follows the constitutive relationship given by (1) can sustain compression unlike a fabric which cannot. The maximum compression stress which can be obtained from (1) is -1 when the strain is (or, strains are) infinetly negative. No negative stresses were obtained for the problem solved in this effort.

Using the data from Figure 2 and experimenting with several linear functions for V_1 and V_2, the following constitutive relation is obtained

$$N_1 = e^{125(\epsilon_1 + 0.5\epsilon_2 - 25\epsilon_1\epsilon_2)} - 1$$
$$N_2 = e^{125(\epsilon_2 + 0.5\epsilon_1 - 25\epsilon_1\epsilon_2)} - 1 \qquad (3)$$

Constant stress profiles for (3) are shown in Figure 3. These contours agree well with those in Figure 2 in the first quadrant and have the shifting and curvature properties similar to the profiles shown in Figure 2. The potential energy can now be easily obtained in terms of the strains.

METHOD OF ANALYSIS. For simplicity we assume that the potential energy is a function of two field variables X_1 and X_2 . Thesefield variables are the strains ϵ_1 and ϵ_2 in the detailed analysis presented in the next section. Extension to the case when more nodal variables are used is straight forward. X_1 and X_2 are approximated by the nodal variables X and Y over an element domain. That is, we have

$$\Pi = F(X_1,X_2) = \text{ the potential energy} \tag{4}$$

where $X_1 = X_1(X,Y)$

$X_2 = X_2(X,Y)$

Then, the gradient of the potential energy is

$$g = \begin{bmatrix} \frac{\partial \Pi}{\partial X} \\ \frac{\partial \Pi}{\partial Y} \end{bmatrix} = \begin{bmatrix} \frac{\partial F}{\partial X_1}\frac{\partial X_1}{\partial X} + \frac{\partial F}{\partial X_2}\frac{\partial X_2}{\partial X} \\ \frac{\partial F}{\partial X_1}\frac{\partial X_1}{\partial Y} + \frac{\partial F}{\partial X_2}\frac{\partial X_2}{\partial Y} \end{bmatrix} \tag{5}$$

or

$$g = \frac{\partial F}{\partial X_1}\begin{bmatrix} \frac{\partial X_1}{\partial X} \\ \frac{\partial X_1}{\partial Y} \end{bmatrix} + \frac{\partial F}{\partial X_2}\begin{bmatrix} \frac{\partial X_2}{\partial X} \\ \frac{\partial X_2}{\partial Y} \end{bmatrix} = \frac{\partial F}{\partial X_1}M_1 + \frac{\partial F}{\partial X_2}M_2 \tag{6}$$

The tangent matrix is written as

$$k = \left[\frac{\partial}{\partial X} g \;\vdots\; \frac{\partial}{\partial Y} g \right] \tag{7}$$

which in this case becomes

$$k = \frac{\partial^2 F}{\partial X_1^2} M_1 M_1^T + \frac{\partial^2 F}{\partial X_1 \partial X_2} \left(M_1 M_2^T + M_2 M_1^T \right) + \frac{\partial^2 F}{\partial X_2^2} M_2 M_2^T$$
$$+ \frac{\partial F}{\partial X_1} J X_1 \quad + \frac{\partial F}{\partial X_2} J X_2 \tag{8}$$

where

$$J Z = \begin{bmatrix} \frac{\partial^2 Z}{\partial X^2} & \frac{\partial^2 Z}{\partial Y \partial Z} \\ \frac{\partial^2 Z}{\partial X \partial Y} & \frac{\partial^2 Z}{\partial Y^2} \end{bmatrix}$$

A solution is defined as a point (X, Y) at which Π is a minimum. The Newton-Raphson method can be used to locate stationary points of Π using the element gradient and tangent matrices as follows

$$u_{i+1} = u_i - K_i^{-1} g_i$$

where $u_i = \begin{bmatrix} X_i \\ Y_i \end{bmatrix}$ represents the i'th vector, etc. (9)

GRADIENT AND TANGENT MATRICES FOR CONSTANT STRAIN ELEMENT. The coordinate system used for the element is shown in Figure 4. In the element coordinate system, the strain-displacement relations can be written as

$$\epsilon_1 = \frac{\partial u^e}{\partial X_e} + \frac{1}{2}\left[\left(\frac{\partial u^e}{\partial X_e}\right)^2 + \left(\frac{\partial v^e}{\partial X_e}\right)^2 + \left(\frac{\partial w^e}{\partial X_e}\right)^2\right]$$

$$\epsilon_2 = \frac{\partial v^e}{\partial Y_e} + \frac{1}{2}\left[\left(\frac{\partial u^e}{\partial Y_e}\right)^2 + \left(\frac{\partial v^e}{\partial Y_e}\right)^2 + \left(\frac{\partial w^e}{\partial Y_e}\right)^2\right] \qquad (10)$$

where (u^e, v^e, w^e) are the displacement components of a material point along the element coordinate directions (X_e, Y_e, Z_e), respectively.

We now introduce the relationships between the element Cartesian coordinates and the area coordinates (see ref. 8 and Figure 4).

$$\xi_1 = (a_1 + b_1 X_e + c_1 Y_e)/(2A)$$

$$\xi_2 = (a_2 + b_2 X_e + c_2 Y_e)/(2A)$$

$$\xi_3 = (a_3 + b_3 X_e + c_3 Y_e)/(2A)$$

$$A = \frac{1}{2}\det\begin{bmatrix} 1 & X_e^1 & Y_e^1 \\ 1 & X_e^2 & Y_e^2 \\ 1 & X_e^3 & Y_e^3 \end{bmatrix} =$$ area of element

$$a_i = X_e^j Y_e^k - X_e^k Y_e^j$$

$$b_i = Y_e^j - Y_e^k$$

$$c_i = X_e^k - X_e^j$$

(X_e^i, Y_e^i) = the Cartesian coordinates of node i

and (i, j, k) is an even permutation of (1, 2, 3).

Next, using the traingular constant-strain element interpolation functions, we interpolate u^e, v^e and w^e as follows

$$\begin{aligned} u^e &= \phi^T u^n \quad , \quad u^n = (u_1^e, u_2^e, u_3^e)^T \quad , \quad \phi^T = (\xi_1, \xi_2, \xi_3) \\ v^e &= \phi^T v^n \quad , \quad v^n = (v_1^e, v_2^e, v_3^e)^T \\ w^e &= \phi^T w^n \quad , \quad w^n = (w_1^e, w_2^e, w_3^e)^T \end{aligned} \tag{11}$$

where (u_i^e, v_i^e, w_i^e) are the nodal displacements of the i'th node. With the definitions given above, the strain in the warp direction can be written as

$$\epsilon_1 = \phi_{,x_e}^T u^n + \frac{1}{2}\left[u^{n^T}\phi_{,x_e}\phi_{,x_e}^T u^n + v^{n^T}\phi_{,x_e}\phi_{,x_e}^T v^n + w^n\phi_{,x_e}\phi_{,x_e}^T w^n\right]$$

But, $\phi_{,x_e}^T = \frac{\partial}{\partial x_e}(\xi_1, \xi_2, \xi_3) = \frac{1}{2A}(b_1, b_2, b_3)$

so the dot products become

$$\begin{aligned} \phi_{,x_e}^T u^n &= \frac{1}{2A}\sum_i b_i u_i = \alpha_b \\ \phi_{,x_e}^T v^n &= \frac{1}{2A}\sum_i b_i v_i = \beta_b \\ \phi_{,x_e}^T w^n &= \frac{1}{2A}\sum_i b_i w_i = \gamma_b \end{aligned} \tag{12}$$

and

$$\epsilon_1 = \alpha_b + \frac{1}{2}(\alpha_b^2 + \beta_b^2 + \gamma_b^2)$$

Similarly $\epsilon_2 = \beta_a + \frac{1}{2}(\alpha_a^2 + \beta_a^2 + \gamma_a^2)$ (13)

The gradient of the element's potential energy is computed as follows

$$g_e = \begin{bmatrix} \frac{\partial \pi_e}{\partial u_1^e} \\ \frac{\partial \pi_e}{\partial v_2^e} \\ \vdots \\ \frac{\partial \pi_e}{\partial w_3^e} \end{bmatrix} \tag{14}$$

where $$\Pi_e = \int_{\Omega_e} F(\epsilon_1,\epsilon_2)\, d\Omega_e - W_e \qquad (15)$$

$F(\epsilon_1,\epsilon_2)$ = the areal energy density functional given by

$$F(\epsilon_1,\epsilon_2) = \frac{1}{2}\,\epsilon_1 N_1(\epsilon_1,\epsilon_2) + \frac{1}{2}\,\epsilon_2 N_2(\epsilon_1,\epsilon_2)$$

W_e = the work done on the element by external forces,

and Ω_e = the domain of the element.

Then, the gradient becomes

$$g = \int_{\Omega_e} \left[\frac{\partial F}{\partial \epsilon_1} \begin{bmatrix} \frac{\partial \epsilon_1}{\partial u_1^e} \\ \frac{\partial \epsilon_1}{\partial v_1^e} \\ \frac{\partial \epsilon_1}{\partial w_1^e} \\ \vdots \\ \frac{\partial \epsilon_1}{\partial w_3^e} \end{bmatrix} + \frac{\partial F}{\partial \epsilon_2} \begin{bmatrix} \frac{\partial \epsilon_2}{\partial u_1^e} \\ \frac{\partial \epsilon_2}{\partial v_1^e} \\ \frac{\partial \epsilon_2}{\partial w_1^e} \\ \vdots \\ \frac{\partial \epsilon_2}{\partial w_3^e} \end{bmatrix} \right] d\Omega_e - \begin{bmatrix} \frac{\partial W_e}{\partial u_1^e} \\ \frac{\partial W_e}{\partial v_1^e} \\ \frac{\partial W_e}{\partial w_1^e} \\ \vdots \\ \frac{\partial W_e}{\partial w_3^e} \end{bmatrix} \qquad (16)$$

or, in a shorter notation

$$g = \int_{\Omega_e} \left[\frac{\partial F}{\partial \epsilon_1} M_1 + \frac{\partial F}{\partial \epsilon_2} M_2 \right] d\Omega_e - \frac{\partial W_e}{\partial u_e^T} \qquad (17)$$

and the element tangent matrix becomes

$$K = \int_{\Omega_e} \left[\frac{\partial^2 F}{\partial \epsilon_1^2} M_1 M_1^T + \frac{\partial^2 F}{\partial \epsilon_1 \partial \epsilon_2} \left[M_1 M_2^T + M_2 M_1^T \right] + \frac{\partial^2 F}{\partial \epsilon_2^2} M_2 M_2^T + \frac{\partial F}{\partial \epsilon_1} J\epsilon_1 + \frac{\partial F}{\partial \epsilon_2} J\epsilon_2 \right] d\Omega_e - \frac{\partial^2 W_e}{\partial u_e \partial u_e^T} \qquad (18)$$

where
$$J\epsilon_i = \begin{bmatrix} \dfrac{\partial^2 \epsilon_i}{\partial u_i^{e\,2}} & \dfrac{\partial^2 \epsilon_i}{\partial v_i^e \partial u_i^e} & \dfrac{\partial^2 \epsilon_i}{\partial w_i^e \partial u_i^e} & \dfrac{\partial^2 \epsilon_i}{\partial u_2^e \partial u_i^e} & \cdots \\ \dfrac{\partial^2 \epsilon_i}{\partial u_i^e \partial v_i^e} & \cdots & & & \\ \cdots & & & & \end{bmatrix} \tag{18}$$

or
$$J\epsilon_i = \left[\begin{array}{c|c|c|c|c} \dfrac{\partial}{\partial u_i^e} M_i & \dfrac{\partial}{\partial v_i^e} M_i & \dfrac{\partial}{\partial w_i^e} M_i & \cdots & \dfrac{\partial}{\partial w_3^e} M_i \end{array} \right]$$

Calculating, we obtain

$$M_1 = \begin{bmatrix} \frac{b_1}{2A}(1 + \alpha_b) \\ b_1 \beta_b \\ b_1 \gamma_b \\ \frac{b_2}{2A}(1 + \alpha_b) \\ b_2 \beta_b \\ b_2 \gamma_b \\ \frac{b_3}{2A}(1 + \alpha_b) \\ b_3 \beta_b \\ b_3 \gamma_b \end{bmatrix} \qquad M_2 = \begin{bmatrix} a_1 \alpha_a \\ \frac{a_1}{2A}(1 + \beta_a) \\ a_1 \gamma_a \\ a_2 \alpha_a \\ \frac{a_2}{2A}(1 + \beta_a) \\ a_2 \gamma_a \\ a_3 \alpha_a \\ \frac{a_3}{2A}(1 + \beta_a) \\ a_3 \gamma_a \end{bmatrix} \tag{19}$$

and the terms in $J\epsilon_1$ and $J\epsilon_2$ are found from

$$\frac{\partial^2\epsilon_1}{\partial u_i^e \partial u_j^e} = \frac{\partial^2\epsilon_1}{\partial v_i^e \partial v_j^e} = \frac{\partial^2\epsilon_1}{\partial w_i^e \partial w_j^e} = \frac{b_i\, b_j}{(2A)^2}$$

$$\frac{\partial^2\epsilon_1}{\partial u_i^e \partial v_j^e} = \frac{\partial^2\epsilon_1}{\partial u_i^e \partial w_j^e} = \frac{\partial^2\epsilon_1}{\partial v_i^e \partial w_j^e} = 0$$

$$\frac{\partial^2\epsilon_2}{\partial u_i^e \partial u_j^e} = \frac{\partial^2\epsilon_2}{\partial v_i^e \partial v_j^e} = \frac{\partial^2\epsilon_2}{\partial w_i^e \partial w_j^e} = \frac{a_i\, a_j}{(2A)^2} \tag{20}$$

$$\frac{\partial^2\epsilon_2}{\partial u_i^e \partial v_j^e} = \frac{\partial^2\epsilon_2}{\partial u_i^e \partial w_j^e} = \frac{\partial^2\epsilon_2}{\partial v_i^e \partial w_j^e} = 0$$

Using the constitutive relation given in (1) we can compute the rate of change of the energy density function in (15) and obtain the coeffecients

$$\frac{\partial F}{\partial\epsilon_1}, \frac{\partial F}{\partial\epsilon_2}, \frac{\partial^2 F}{\partial\epsilon_1^2}, \frac{\partial^2 F}{\partial\epsilon_1\partial\epsilon_2} \text{ and } \frac{\partial^2 F}{\partial\epsilon_2^2}$$

Now with a description of a fabric's geometry, boundary conditions and loading, solutions can be obtained using the algorithm described above.

Uniformly Loaded Square Membrane. One quarter of a uniformly loaded square fabric membrane is shown with boundary conditions in Figure 5. To obtain a solution to this problem the following approximate expression for the work done was used.

$$W = \int q\, d\text{Volume} = \int_\Omega q\, z\, dx\, dy = q \int_\Omega w\, dx\, dy \tag{21}$$

where q = the pressure applied to the membrane.

w = the vertical displacement of the membrane.

and Ω = the (x,y) domain of the fabric.

For the constant strain element described above we find (for an element of Ω).

$$\frac{\partial W_e}{\partial u_e^T} = \frac{qA}{3}\begin{bmatrix}0\\0\\1\\0\\0\\1\\0\\0\\1\end{bmatrix} \qquad (22)$$

and $$\frac{\partial^2 W_e}{\partial u_e\,\partial u_e^T} = \left[\,0\,\right]$$

where $u_e^T = (u_1^e, v_1^e, w_1^e, u_2^e, v_2^e, w_2^e, u_3^e, v_3^e, w_3^e)^T$ and W_e in the portion of the external work done on the elements.

The deflections, strains and stresses for a uniformly loaded square membrane with the boundary conditions given in Figure 5 and a constitutive relation given by (3) were determined. The size of the membrane was 10 in X 10 in, (represents one quarter of a 20 in X 20 in uniformly loaded membrane). Figure 6 shows the maximum displacement and load shown indicates that when the membrane is nearly flat it is very flexible but after it has been deformed it becomes stiff. The values of the membrane stress, N_x, are shown in Figure 7 for a pressure of 0.5 lb/in^2. The distribution for N_y was entirely symmetrical with respect to N_x and is not shown. This data indicates that the stresses are very low in the corners of the uniformly loaded square membrane and are the largest along the centerlines. Corresponding values of the strains E_x and E_y are shown in Fig. 8. The strains indicate that the fabric will tend to pull tight across the two centerlines of the square but remain relatively unloaded near the corners.

The effect of mesh size on the accuracy of the solution was determined by using Richardson's extrapolation method to study the accuracy of the displacements. This was accomplished by assuming the center deflection of the square membrane was related to the mesh size as follows.

$$w(h) = w(o) + Ch^P$$

where w(h) is the center deflection for mesh size h and C,P are constants. Solutions were obtained at a pressure of $q = 10^{-4}$ lb/in^2. for 6 X 6, 8 X 8 and 10 X 10 meshes (i.e. for h = 1/6, 1/8, 1/10). The values for h and w(h) were used to determine the convergence rate (P = 1.69) shown in Figure 9.

Conclusions. The elastic nonlinear behavior of fabrics can be modeled by exponential functions in which stresses are determined as a function of the strains. These stress-strain relationships allow the stresses in the deformed fabrics to be determined by a nonlinear displacement finite element method. This approach is useful for determining the deformations, strains and stresses in uniformly loaded square fabric membranes. The approach here can be modified and used for designing tents and inflatable fabric structures.

References:

1. R. E. Sebring and W. D. Freeston, "Biaxial Tensile Tester for Fabrics", USA NLABS TR67-71-GP (NTIS AD-658684), May 1967.

2. P.J. Remington, J. C. O'Callahan and R. Madden, "Analysis of Stresses and Deflections in Frame Supported Tents," USA NLABS TR 75-31, April 1974.

3. E. C. Steeves, "Mathematical Modeling of the Stress-Strain Behavior of Fabrics," USA NLABS TR82-009 (NTIS AD-A115058), March 1982.

4. J. Christoffersen, "Fabrics: Orthotropic Materials with a Stress-Free Shear Mode," JAM, Vol 47, March 1980 pp71-74.

5. H. Minami and Y. Nakahara, "An Application of Finite Element Method To The Deformation Analysis of Coated Plain-Weave Fabrics," J. of Coated Fabrics, Vol 11, April 1981, pp 310-327.

6. A. R. Johnson, "Finite Element Analysis of Axisymmetric Rubber Membranes, "PhD Dissertation, Boston University, 1981.

7. A. R. Johnson, "Large Deformations and Stability of Axisymmetric Mooney Membranes - Finite Element Solutions," Trans. of the Twenty-Eighth Conference of Army Mathematicians, U.S. Army Research Office Report No. 83-1, February 1983.

8. O.C. Zienkiewicz, The Finite Element Method, McGraw-Hill Book Co., 1977.

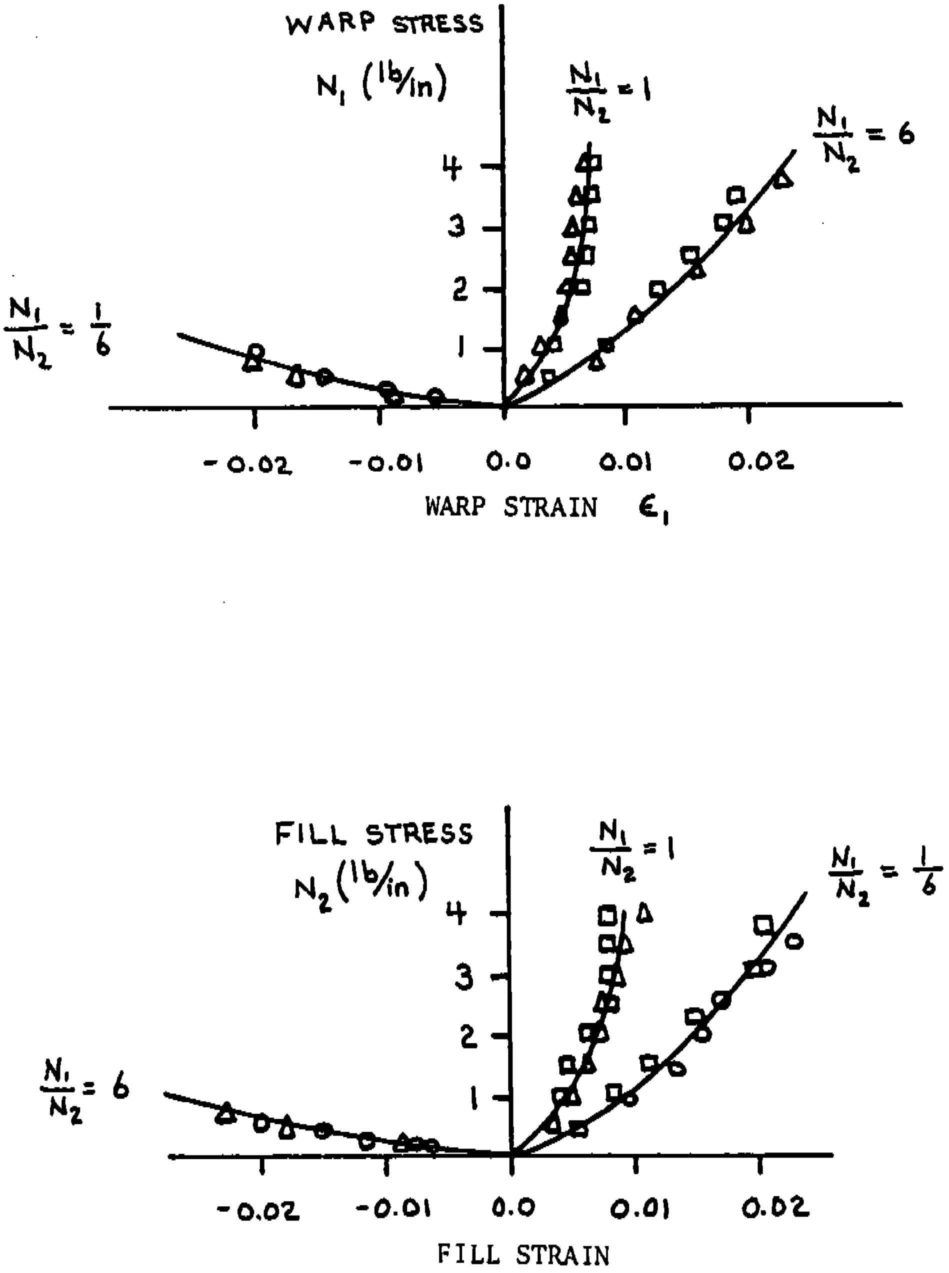


FIGURE 1. Biaxial fabric test data for 2.6 oz/yd^2 cotton typerwriter ribbon cloth. (taken from Ref. 2).

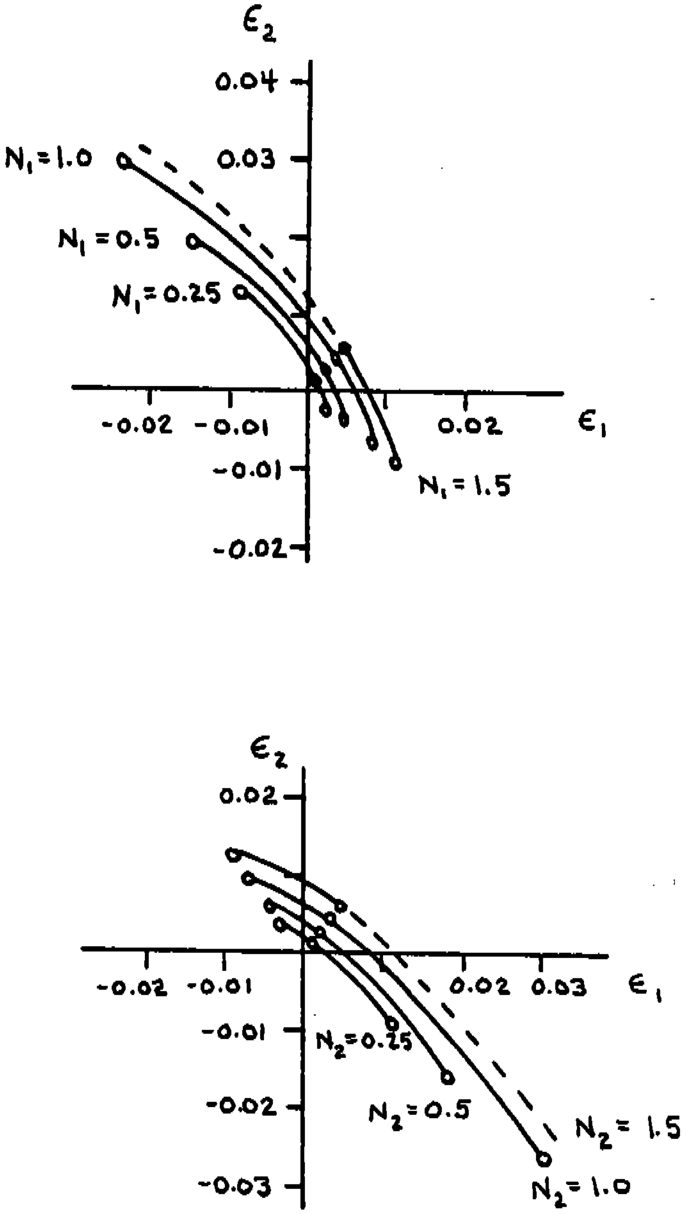


FIGURE 2. Contours of constant stress as a function of ϵ_1 and ϵ_2 for 2.6 OZ/yd^2 cotton typerwriter ribbon cloth.

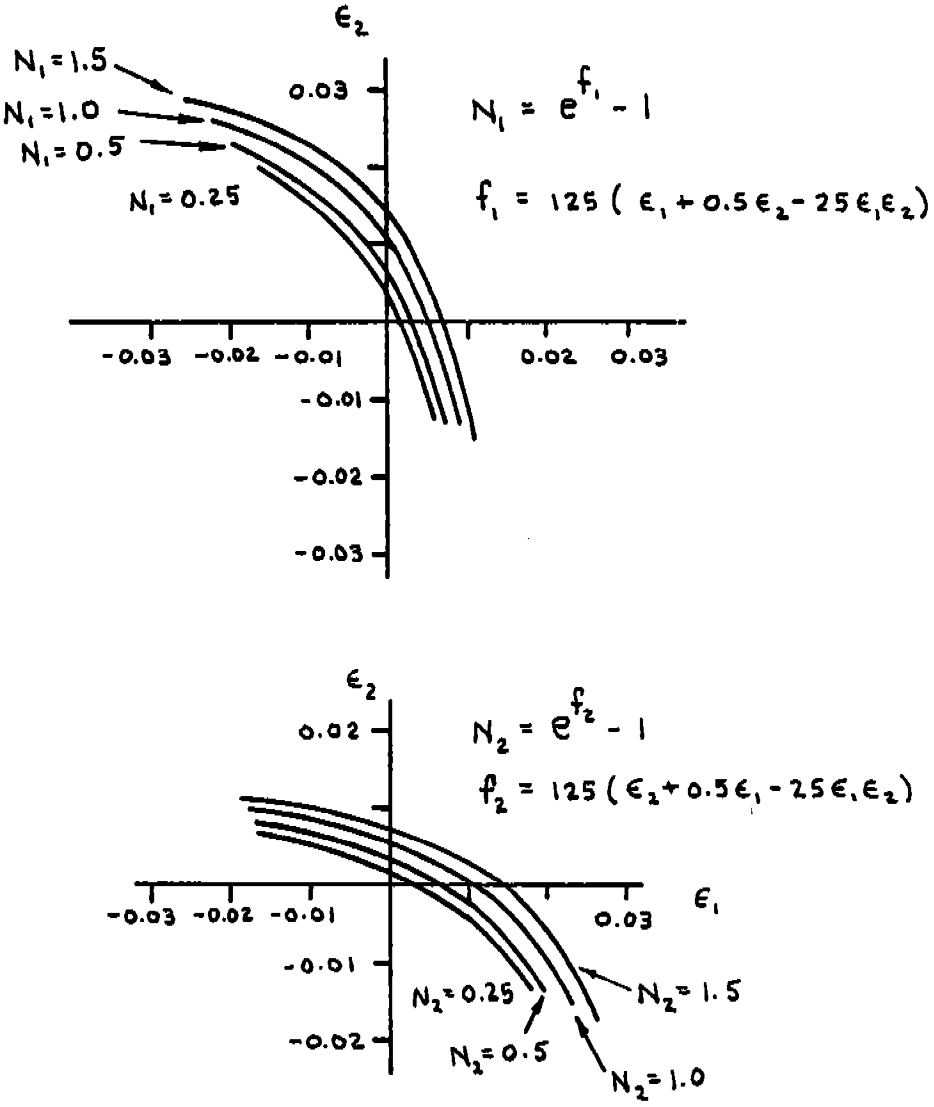


Figure 3. Contours of constant stress as function of ϵ_1 and ϵ_2 for exponential constitutive law.

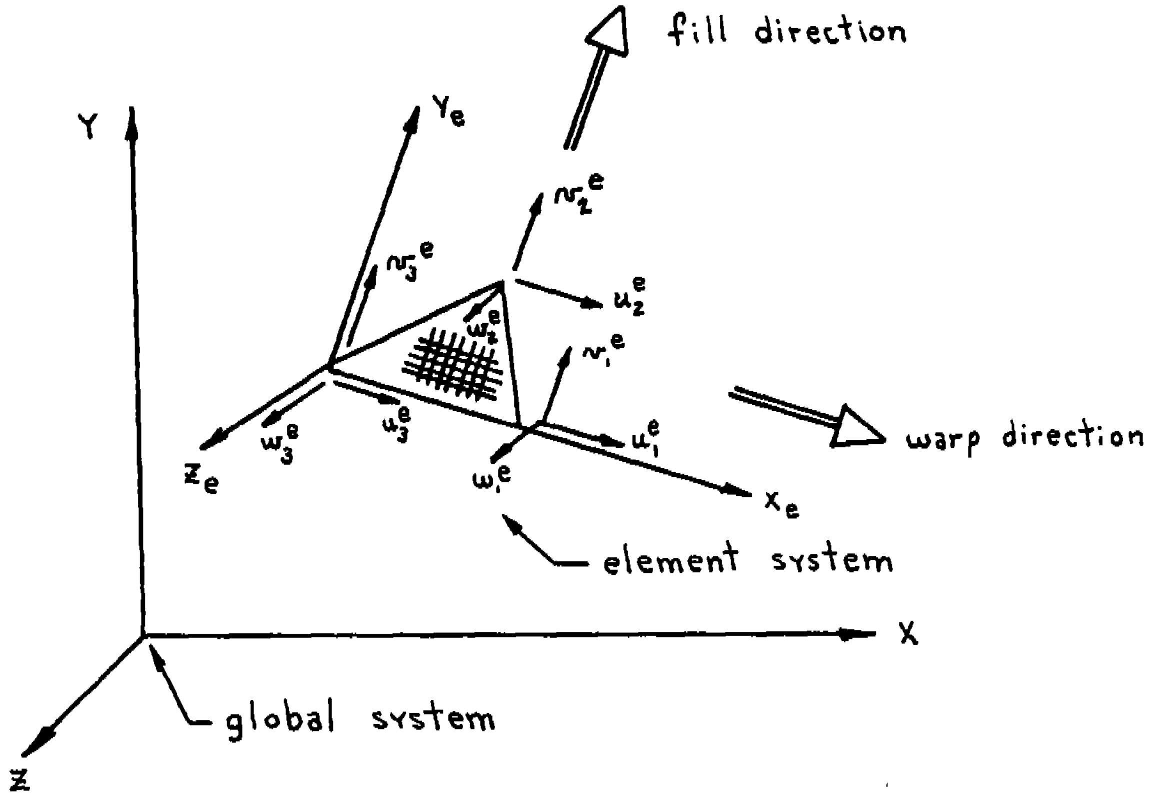


FIGURE 4. Coordinate System For The Fabric Element.

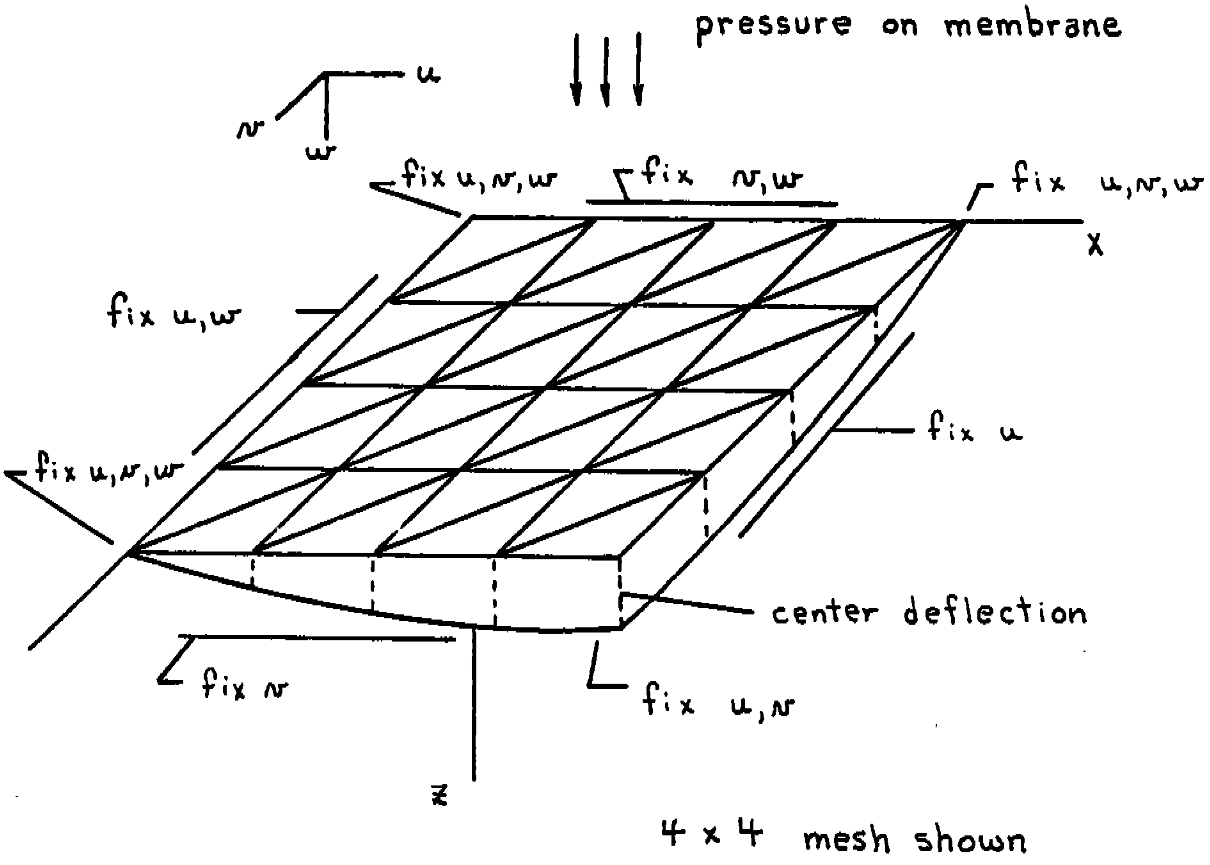


FIGURE 5. Loading, boundary conditions, and typical mesh for flat square fabric membrane.

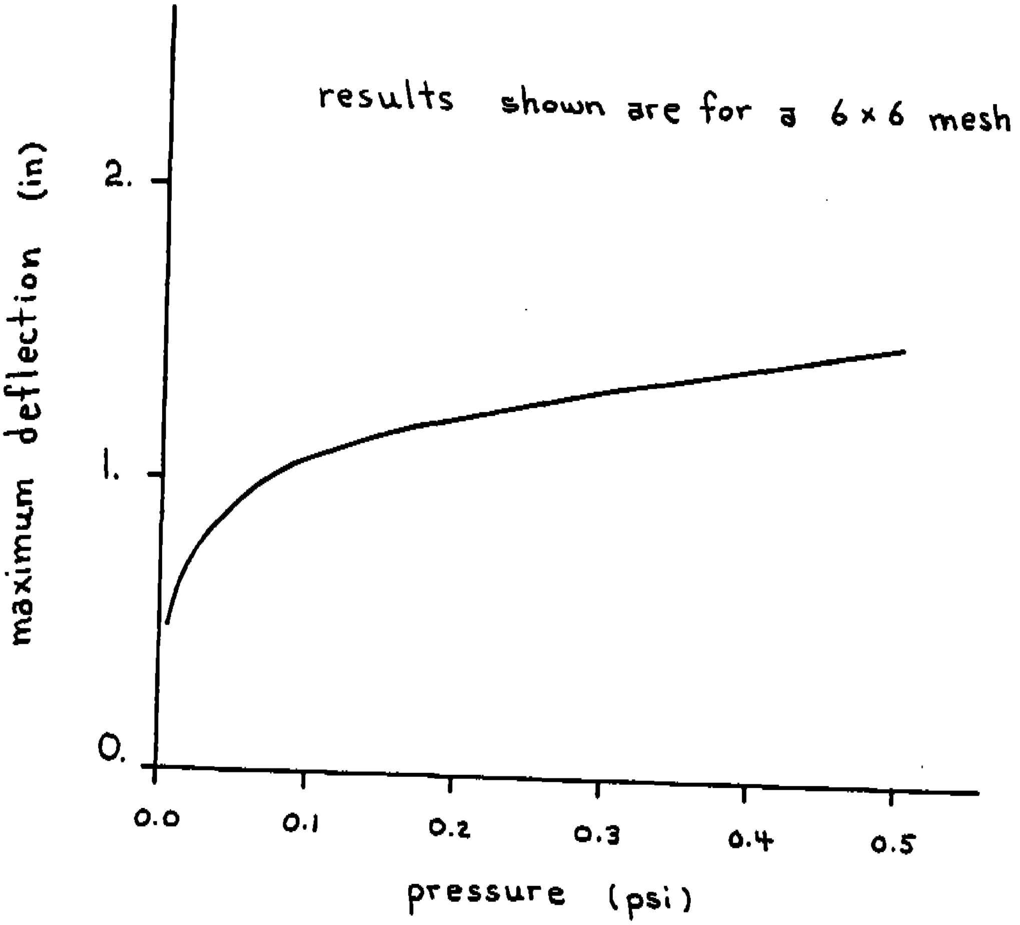


FIGURE 6. Maximum deflection vs pressure for an initially flat square fabric membrane.

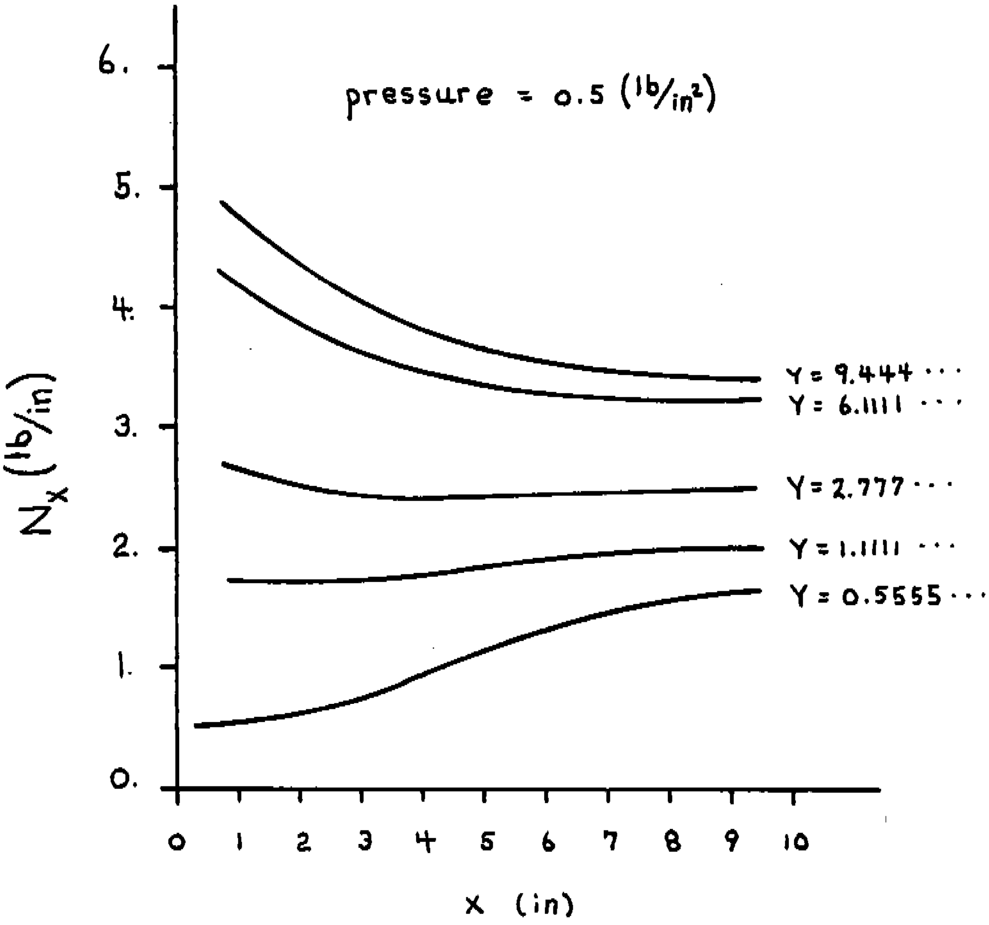


FIGURE 7. Membrane stress N_x vs X for a 6 X 6 mesh (stress taken at geometric center of elements).

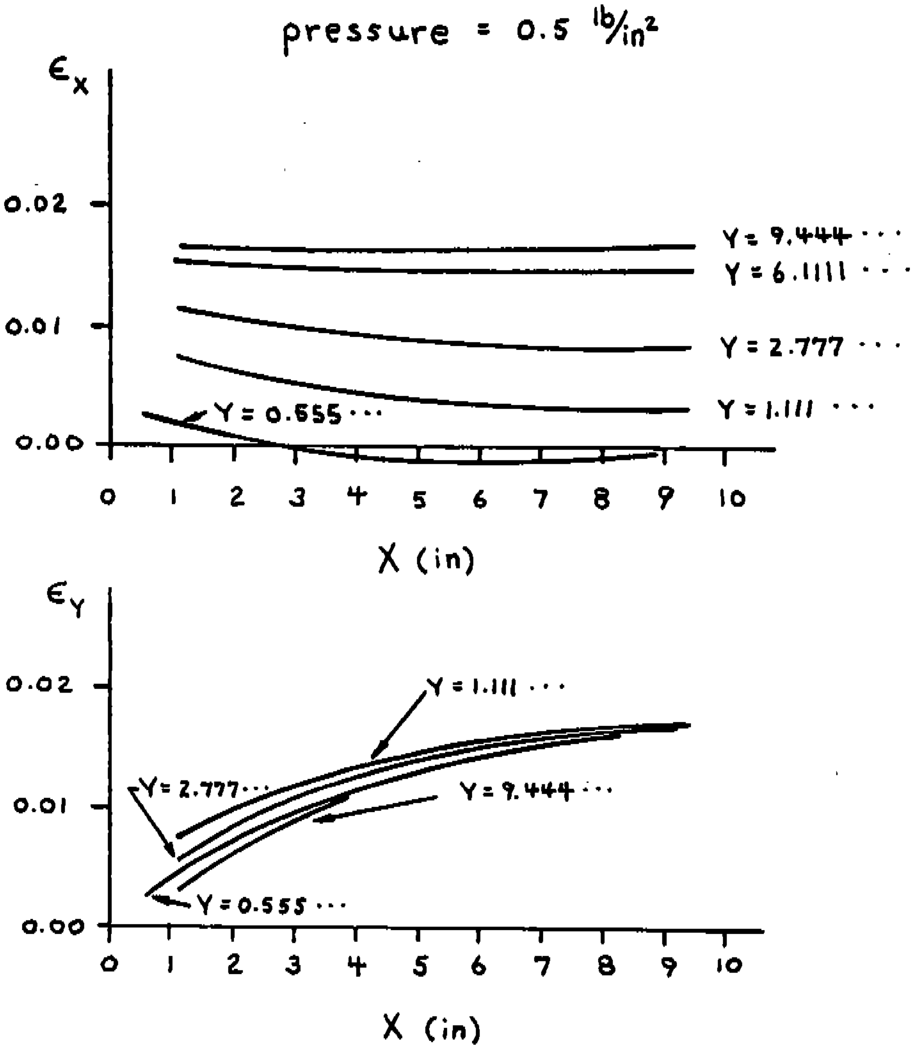


FIGURE 8. Membrane strains for a 6 X 6 mesh.

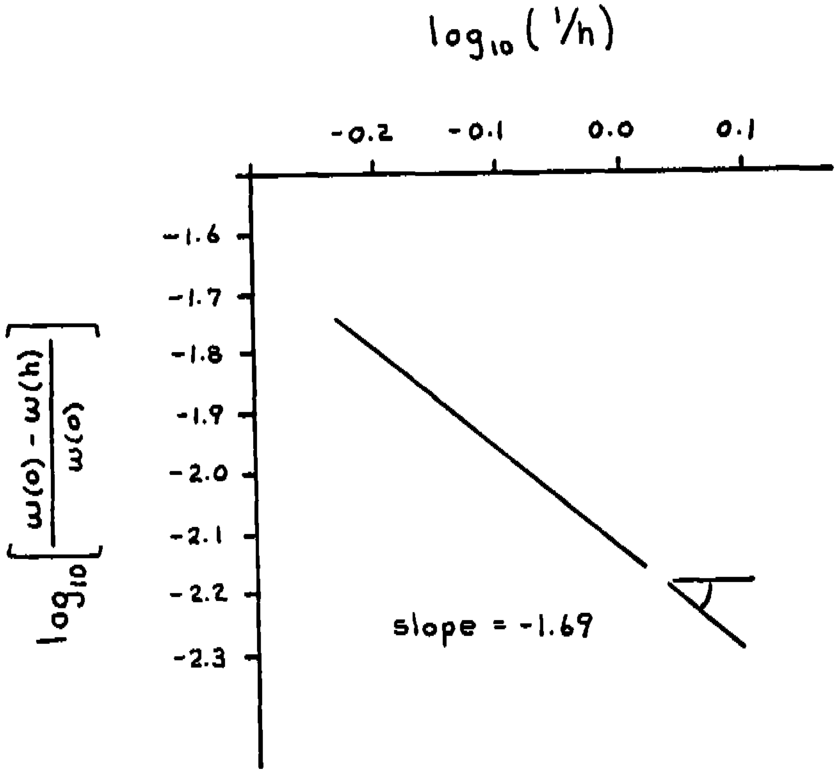


FIGURE 9. Convergence of center deflection with respect to mesh size for a small pressure (P = 0.0001 lb/in^2).

EXPLICIT FORMULAS FOR C^n PIECEWISE HERMITE BASIS FUNCTIONS

Royce W. Soanes, Jr.
US Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. Completely factored forms of the piecewise Hermite basis functions will be derived. All necessary coefficients for any level of smoothness will be shown to reside conveniently in Pascal's triangle.

I. INTRODUCTION. We begin with a theoretical characterization of the C^n basis functions with which we are dealing. For a given node x_i in R, there is a basis function $H_{ij}(x)$ associated with the jth derivative of any function f at x_i, where j ranges from o to n. In addition, each basis function is nonzero on only two adjacent subintervals. Continuing with the definition of the H's, we wish an approximation F to f of the form:

$$F(x) = \sum_{j=0}^{n} H_{ij}(x)f^{(j)}(x_i) + H_{i+1j}(x)f^{(j)}(x_{i+1})$$

where $x_i \leq x \leq x_{i+1}$.

In order that F and its n derivatives will agree with f and n of its derivatives at nodes x_i and x_{i+1}, it is sufficient that the basis functions associated with arbitrary node i obey the following conditions:

$$H_{ij}^{(k)}(x_i) = \delta_{jk}$$

and

$$H_{ij}^{(k)}(x_{i-1}) = H_{ij}^{(k)}(x_{i+1}) = 0$$

where $0 \leq j,k \leq n$.

On a given subinterval, therefore, each H must obey n+1 conditions on the left extreme and n+1 conditions on the right extreme. The H's may therefore be represented by two distinct polynomials of degree 2n+1.

The following series of pictures depicts the C^4 basis functions (scaled) and their derivatives. The five functions across the top are the basis functions associated with the 0th through the 4th derivatives of f and the functions underneath them are their successive derivatives. Note that the functions along the diagonal are nonzero in the center while all off diagonal functions are zero there.

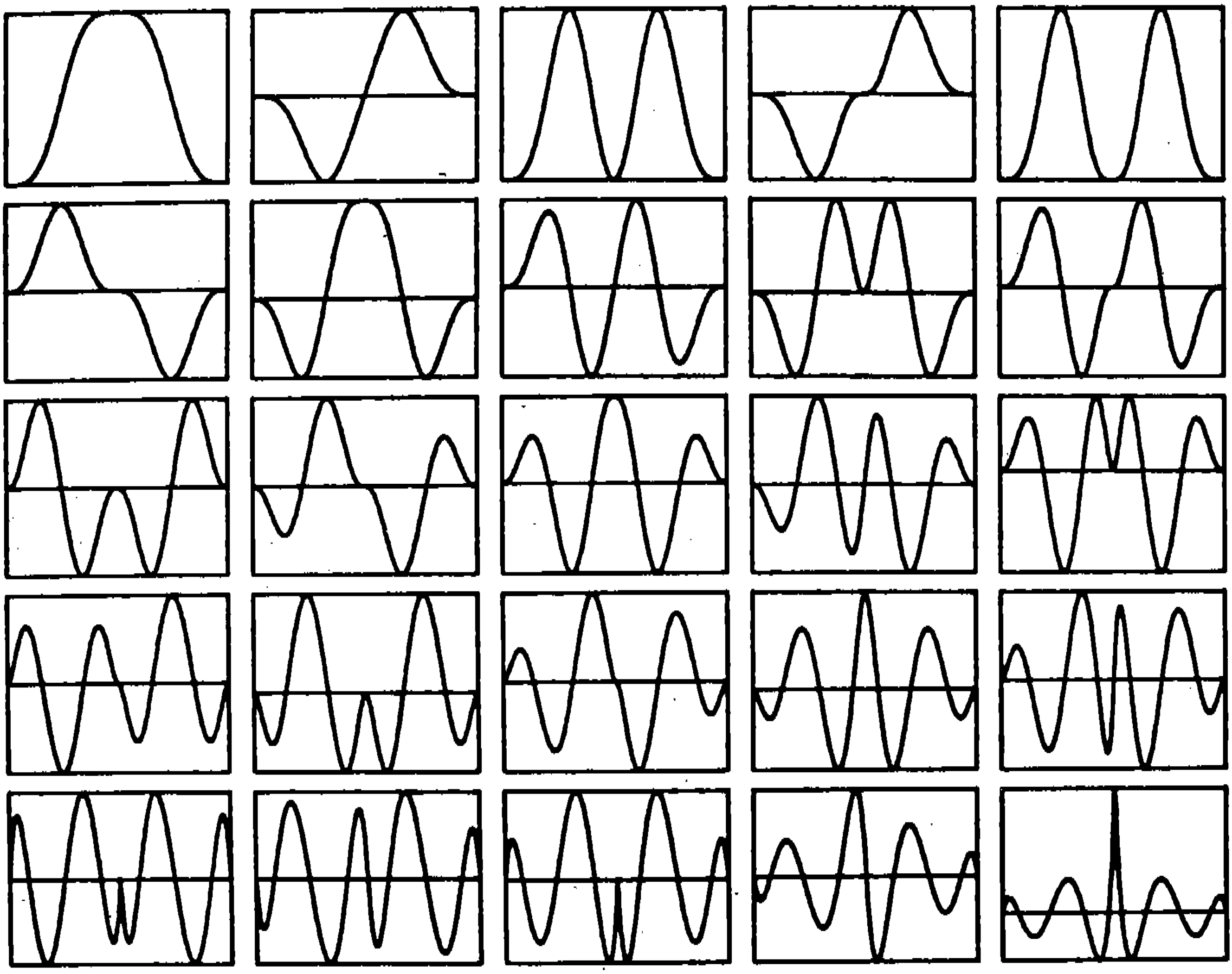


Figure 1. C^4 Basis Functions and Their Derivatives.

Although subsequent analysis will enable us to compute all the basis functions for any given level of smoothness (n) in an extremely simple way, the author has not seen anything similar mentioned or referenced in any finite element text thus far.

II. DERIVATION. We begin by defining a finite support Taylor series (FSTS). Take the ordinary Taylor series for f around node x_i, truncate it beyond nth derivative terms, and multiply each term by a function which will have the effect of (1) not disturbing the truncated Taylor series at all at node x_i and (2) zeroing the series and n of its derivatives at nodes x_{i-1} and x_{i+1}. If we do this for each node x_i, calling the result $F_i(x)$, we get a global approximation to f which agrees with the truncated Taylor series of f at each node by simply summing the F_i. This is just an alternative way of defining the piecewise Hermite approximation which we will find quite useful.

$$\text{TS:} \quad f(x) = \sum_{j=0}^{\infty} f^{(j)}(x_i)(x-x_i)^j/j! \tag{1}$$

$$\text{FSTS:} \quad F_i(x) = \sum_{j=0}^{n} f^{(j)}(x_i)(x-x_i)^j g_j(R_i(x))/j! \tag{2}$$

where

$$\begin{aligned} R_i(x) &= 1 \quad \text{if } x < x_{i-1} \text{ or } x > x_{i+1} \\ &= (x_i-x)/(x_i-x_{i-1}) \text{ if } x_{i-1} \le x < x_i \\ &= (x-x_i)/(x_{i+1}-x_i) \text{ if } x_i \le x \le x_{i+1} \end{aligned} \tag{3}$$

$R_i(x)$ is just one minus the hat function associated with node i or just the relative position of x in either the left or the right hand subinterval. The domain of the g functions is therefore just the interval [0,1].

The objective now is to determine the g's. Since we want F_i and its derivatives to behave in a certain manner, we must first differentiate $F_i(x)$ an arbitrary number of times. Using Leibniz's rule for differentiating a product, we have:

$$F_i^{(m)}(x) = \sum_{j=0}^{n} f^{(j)}(x_i) \sum_{k=0}^{\min\{j,m\}} \binom{m}{k}(x-x_i)^{j-k} g_j^{(m-k)}(R_i(x))(R_i'(x))^{m-k}/(j-k)! \tag{4}$$

and substituting $x = x_i$ in Eq. (4) we have

$$F_i^{(m)}(x_i) = \sum_{j=0}^{m} f^{(j)}(x_i)\binom{m}{j} g_j^{(m-j)}(0)(R_i'(0))^{m-j} \tag{5}$$

we may define $R_i'(0)$ to be $R_i'(\pm\varepsilon)$ or take limits from either side of x_i.

We now want conditions on the g's which are sufficient for:

$$F_i^{(m)}(x_i) = f^{(m)}(x_i) \tag{6}$$

and

$$F_i^{(m)}(x_{i-1}) = F_i^{(m)}(x_{i+1}) = 0 \tag{7}$$

for $0 \le m \le n$.

We may glean these conditions from Eqs. (4) and (5). Conditions on the g's sufficient for Eqs. (6) and (7) may be seen to be:

$$g_j(0) = 1 \qquad 0 \leq j \leq n \tag{8}$$

$$g_j^{(m)}(0) = 0 \qquad 0 \leq j < n,\ 1 \leq m \leq n-j \tag{9}$$

$$g_j^{(m)}(1) = 0 \qquad 0 \leq j \leq n,\ 0 \leq m \leq n \tag{10}$$

The 2n-j+2 conditions on g_j may therefore be met by a polynomial of degree 2n-j+1; the product of g_j and $(x-x_i)^j$ in Eq. (2) is therefore of degree 2n+1 for all j, as expected.

The g of lowest degree is therefore g_n, which has defining conditions:

$$g_n(0) = 1$$

and

$$g_n^{(m)}(1) = 0 \qquad 0 \leq m \leq n$$

This g may be obtained by inspection and is:

$$g_n(x) = (1-x)^{n+1} \tag{11}$$

Now, from Eq. (10), we may observe that all the g's have the same derivative behavior at x=1. We therefore need only define $g_j(x)$ as the product of some unknown polynomial $h_j(x)$ and $g_n(x)$:

$$g_j(x) = h_j(x)(1-x)^{n+1} \qquad 0 \leq j \leq n \tag{12}$$

where h_j is a polynomial of degree n-j:

$$h_j(x) = \sum_{k=0}^{n-j} a_k x^k \qquad 0 \leq j \leq n \tag{13}$$

It may seem at first glance that the a's should have an extra subscript - namely j, since we are seeking n+1 sets of coefficients. As will become immediately apparent, however, one set is sufficient and all other sets are subsets of this one. This subset property and the fact that the largest set of a's may be obtained in an almost trivial manner, is what makes the result of this analysis truly simple indeed.

If we now obtain the mth derivative of $g_j(x)$ and evaluate it at x = 0, we get:

$$g_j^{(m)}(0) = m! \sum_{k=0}^{m} a_k (-1)^{m-k} \binom{n+1}{m-k} \tag{14}$$

Using Eqs. (8) and (9), we have:

$$a_o = 1$$

and

$$\sum_{k=0}^{m} a_k(-1)^{m-k}\binom{n+1}{m-k} = 0 \quad \text{for} \quad 1 \leq m \leq n-j \tag{15}$$

This is a lower triangular system, which may be easily solved for the a's by forward substitution. Note that the coefficients do not depend on j, so we might as well solve the largest system (j=0) and obtain <u>all</u> the a's, although only the first n-j+1 a's are needed for $h_j(x)$.

Solving system (15) for j=0 and a few values of n gives us the following table of a's

TABLE 1. COEFFICIENTS OF h_o

n	a_o	a_1	a_2	a_3	a_4	a_5
0	1					
1	1	2				
2	1	3	6			
3	1	4	10	20		
4	1	5	15	35	70	
5	1	6	21	56	126	252

Inspection of this table gives us a very simple recursion for the n set of coefficients in terms of the n-1 set:

$$\overset{n}{a_o} = 1$$

$$\overset{n}{a_k} = \overset{n}{a_{k-1}} + \overset{n-1}{a_k} \qquad 1 \leq k \leq n-1$$

$$\overset{n}{a_n} = 2\overset{n}{a_{n-1}} \tag{16}$$

where the superscripts denote the level of smoothness.

We therefore have here nothing more than one half of Pascal's triangle, viewed at an angle!

Recalling that:

$$g_j(x) = h_j(x)(1-x)^{n+1}$$

and from the FSTS that:

$$H_{ij}(x) = g_j(R_i(x))(x-x_i)^j/j!$$

we have, explicitly:

$$H_{ij}(x) = h_j(R_i(x))(1-R_i(x))^{n+1}(x-x_i)^j/j! \qquad (17)$$

therefore, f may be approximated on $[x_i,x_{i+1}]$ by:

$$\sum_{j=0}^{n} \{f^{(j)}(x_i)h_j\left(\frac{x-x_i}{x_{i+1}-x_i}\right)\left(\frac{x_{i+1}-x}{x_{i+1}-x_i}\right)^{n+1}(x-x_i)^j$$

$$+ f^{(j)}(x_{i+1})h_j\left(\frac{x_{i+1}-x}{x_{i+1}-x_i}\right)\left(\frac{x-x_i}{x_{i+1}-x_i}\right)^{n+1}(x-x_{i+1})^j\}/j! \qquad (18)$$

In order to evaluate the derivatives of the H's, one may expand the polynomials involved and multiply out or one may apply Leibniz's rule a couple of times. The latter course is deemed simpler and more numerically stable since it leaves a result which is in "nearly" fully factored form. The latter method was used to produce Figure 1.

BIVARIATE QUADRATIC SPLINES ON CRISSCROSS TRIANGULATIONS

Charles K. Chui
Center for Approximation Theory
Department of Mathematics
Texas A&M University
College Station, Texas 77843

ABSTRACT. A bivariate C^1 quadratic B-spline basis for the space of C^1 piecewise polynomials with total degree two on a crisscross triangulation is given. This basis has very important algebraic, geometric and approximatic properties, and can be used in a variety of applications. In particular, it can be used adaptively in pattern recognition, image processing and data reduction. In image restoration, for example, it gives much better pictures than the tensor product splines using the same discrete data.

1. INTRODUCTION.

Let D be a domain in R^2 and Δ a grid of straight line segments that partition D into cells. The collection of all functions s in $C^\mu(D)$ such that the restrictions of s to each cell are polynomials with total degree at most k, that is

$$\sum_{0 \le i+j \le k} a_{ij} x^i y^j ,$$

where μ and k are nonnegative integers, is called the space of bivariate splines with (total) degree k in $C^\mu(D)$ on the grid partition Δ, and will be denoted by $S_k^\mu(D,\Delta)$. Clearly, $S_k^\mu(D,\Delta)$ becomes the trivial space of polynomials with total degree k if $\mu \geq k$. More generally, we could allow Δ to consist of algebraic curves. Some fundamental tools have recently been developed in [3] to study these spaces. These methods could be used to determine dimensions, find locally supported splines, etc. as in [5] and [4]. Locally supported splines with minimum supports are most important both in theory and applications. These functions are generalizations of univariate B-splines, and are sometimes called bivariate B-splines, although this name is used by some authors to include all locally supported splines which are strictly positive inside the supports. When Δ is a regular grid partition, these functions can usually be obtained by projections as discussed in [1]. For a fairly up-to-date review of the subject of multivariate splines, the reader is referred to the 95-page survey article [8].

2. B-SPLINES ON CRISSCROSS TRIANGULATIONS.

In application, such as in surface fitting, pattern recognition, image processing, data reduction, etc., D is usually a rectangular region, and depending on the density and variation of the data to be studied, D is divided into rectangular subregions of different sizes. Let

$$a = x_0 < x_1 < \dots < x_{m+1} = b, \quad \text{and}$$
$$c = y_0 < y_1 < \dots < y_{n+1} = d.$$

Two common ways to triangulate the rectangular subregions $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$, $i = 0,\dots,m$ and $j = 0,\dots,n$, are to draw in diagonals with only positive (or only negative) slopes, or to draw in diagonals with both positive and negative slopes of these subregions. They are called uni-diagonal (or type-1) and crisscross (or type-2) triangulations. If, in particular, the partition is regular, i.e. $x_{i+1} - x_i = x_i - x_{i-1}$, $i = 1,\dots,m$, and $y_{j+1} - y_j = y_j - y_{j-1}$, $j = 1,\dots,n$, then a unidiagonal triangulation becomes a three-direction mesh and a crisscross triangulation becomes a four-direction mesh. Very interesting results on a three-direction mesh can be found in [2] and [6]. It has been pointed out in [7], however, that a minimum support for splines in $S_3^1(D,\Delta)$ where Δ is a three-direction mesh, no longer supports a nontrivial B-spline if Δ becomes irregular. Since irregular (i.e. non-uniform) subdivisions are very important in applications, especially in adaptive procedures, one would wish to obtain bivariate B-splines that change continuously with the grid lines. On crisscross triangulations, a cubic C^1 B-spline is given in [4] and a quadratic C^1 B-spline, with slightly larger support, is given in [7]. These B-splines have minimum supports and are continuous in the grid lines $x = x_i$ and $y = y_j$. Since lower degree splines are more desirable, we only discuss the quadratic one.

To give an expression of the B-spline, we give each of the polynomial pieces. A bivariate quadratic polynomial has six coefficients and they are uniquely determined by the six values of the polynomial at the three vertices and the mid-points of the three sides of a triangle. Let A, B and C be the vertices of a triangle and A', B' and C' the mid-points of the sides opposite to A, B and C respectively. Let us use barycentric coordinates; that is, let a, b and c be linear polynomials such that $a(A) = 1$, $b(B) = 1$, $c(C) = 1$, and the values of a, b, c at the other vertices are zero. Then the quadratic polynomial which takes on the values $f(A)$, $f(B)$, $f(C)$, $f(A')$, $f(B')$, $f(C')$ at the six points A, B, C, A', B', C', respectively, is given by

$$\begin{aligned} p(a,b,c) &= a(2a-1)f(A) + b(2b-1)f(B) + c(2c-1)f(C) \\ &\quad + 4[bcf(A') + caf(B') + abf(C')] . \end{aligned}$$

Note that $a + b + c = 1$, so that p is actually a polynomial in two variables. In writing a computer program to give p as a function of x and y, one could simply consider $a = 0$, $b = 0$, $c = 0$ as the (linear) equations of the sides opposite to A, B, C respectively, so normalized that $a(A) = b(B) = c(C) = 1$.

We now give a representation of our C^1 bivariate quadratic spline B_{ij} whose support is an octagon contained in the rectangle $[x_{i-1}, x_{i+2}] \times [y_{j-1}, y_{j+2}]$ as in Fig. 1. Here, the values of B_{ij} at six (appropriate) points of each triangle are given. Of course, since B_{ij} vanishes on the boundary of its octagonal support, we do not have to give its values there. The values A_i, A'_{i+1}, B_j, and B'_{j+1} in Fig. 1 are defined by

$$A_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}, \qquad A'_{i+1} = \frac{x_{i+2} - x_{i+1}}{x_{i+2} - x_i},$$

$$B_j = \frac{y_j - y_{j-1}}{y_{j-1} - y_{j-1}}, \qquad B'_{j+1} = \frac{y_{j+2} - y_{j+1}}{y_{j+2} - y_j}$$

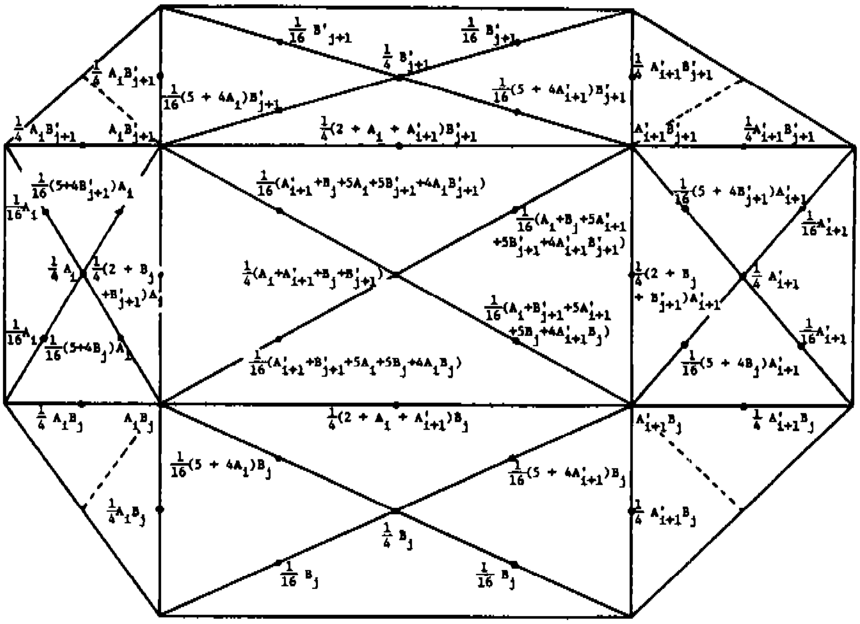

$\frac{1}{16} B'_{j+1}$
$\frac{1}{4} B'_{j+1}$
$\frac{1}{16} B'_{j+1}$
$\frac{1}{4} A_i B'_{j+1}$
$\frac{1}{16}(5 + 4A_i)B'_{j+1}$
$\frac{1}{16}(5 + 4A'_{i+1})B'_{j+1}$
$\frac{1}{4} A'_{i+1} B'_{j+1}$
$\frac{1}{4} A_i B'_{j+1}$
$A_i B'_{j+1}$
$\frac{1}{4}(2 + A_i + A'_{i+1})B'_{j+1}$
$A'_{i+1} B'_{j+1}$
$\frac{1}{4} A'_{i+1} B'_{j+1}$
$\frac{1}{16}(5+4B'_{j+1})A_i$
$\frac{1}{16} A_i$
$\frac{1}{16}(A'_{i+1}+B_j+5A_i+5B'_{j+1}+4A_i B'_{j+1})$
$\frac{1}{16}(A_i+B_j+5A'_{i+1}$
$+5B'_{j+1}+4A'_{i+1}B'_{j+1})$
$\frac{1}{16}(5 + 4B'_{j+1})A'_{i+1}$
$\frac{1}{16} A'_{i+1}$
$\frac{1}{4} A_i$
$\frac{1}{4}(2 + B_j$
$+B'_{j+1})A_i$
$\frac{1}{4}(A_i+A'_{i+1}+B_j+B'_{j+1})$
$\frac{1}{4}(2 + B_j$
$+ B'_{j+1})A'_{i+1}$
$\frac{1}{4} A'_{i+1}$
$\frac{1}{16} A_i$
$\frac{1}{16}(5+4B_j)A_i$
$\frac{1}{16}(A_i+B'_{j+1}+5A'_{i+1}$
$+5B_j+4A'_{i+1}B_j)$
$\frac{1}{16} A'_{i+1}$
$\frac{1}{16}(A'_{i+1}+B'_{j+1}+5A_i+5B_j+4A_i B_j)$
$\frac{1}{16}(5 + 4B_j)A'_{i+1}$
$\frac{1}{4} A_i B_j$
$A_i B_j$
$\frac{1}{4}(2 + A_i + A'_{i+1})B_j$
$A'_{i+1} B_j$
$\frac{1}{4} A'_{i+1} B_j$
$\frac{1}{16}(5 + 4A_i)B_j$
$\frac{1}{16}(5 + 4A'_{i+1})B_j$
$\frac{1}{4} A_i B_j$
$\frac{1}{4} A'_{i+1} B_j$
$\frac{1}{4} B_j$
$\frac{1}{16} B_j$
$\frac{1}{16} B_j$

Fig. 1

Note that these values have very interesting geometric meaning. In addition, they are all positive numbers and

$$A_i + A'_i = B_j + B'_j = 1$$

for all i and j. In Fig. 2, we give a three-dimensional picture of B_{ij} with $x_{i-1}, \ldots, x_{i+2} = 0, 1, 4, 8$ and $y_{j-1}, \ldots, y_{j+2} = 0,3,5,6$.

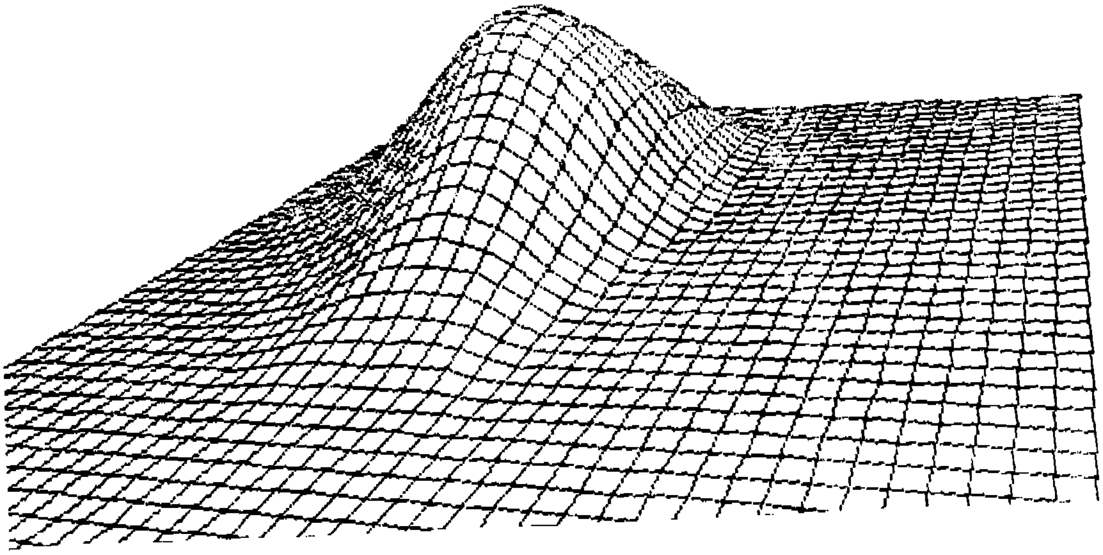


Fig. 2

View: (15,10,1)
Target: (.5,5,.5)
Field: 150°

3. PROPERTIES OF B_{ij}. It is clear from Fig. 1 that B_{ij} is strictly positive inside its support. From Fig. 2, it can also be seen that each (vertical) cross-section of B_{ij} is a bell-shaped curve as expected. A very important feature for approximation is that the B_{ij}'s form a partition of unity; that is,

$$\sum_{j=-1}^{n+1} \sum_{i=-1}^{m+1} B_{ij}(x,y) = 1$$

for all (x,y) in the rectangle $[a,b] \times [c,d]$. Here, we have assigned arbitrary values $x_{-2} \leq x_{-1} \leq x_0 = a < b = x_{m+1} \leq x_{m+2} \leq x_{m+3}$ and $y_{-2} \leq y_{-1} \leq y_0 = c < d = y_{n+1} \leq y_{n+2} \leq y_{n+3}$. Coalescence of x_i's and y_j's are allowed as in the univariate setting. Of course, if $x_i = x_{i+1}$ then the joining condition along this edge becomes C^0 and if $x_i = x_{i+1} = x_{i+2}$ then B_{ij} has a jump on this edge. See Fig. 3a,b.

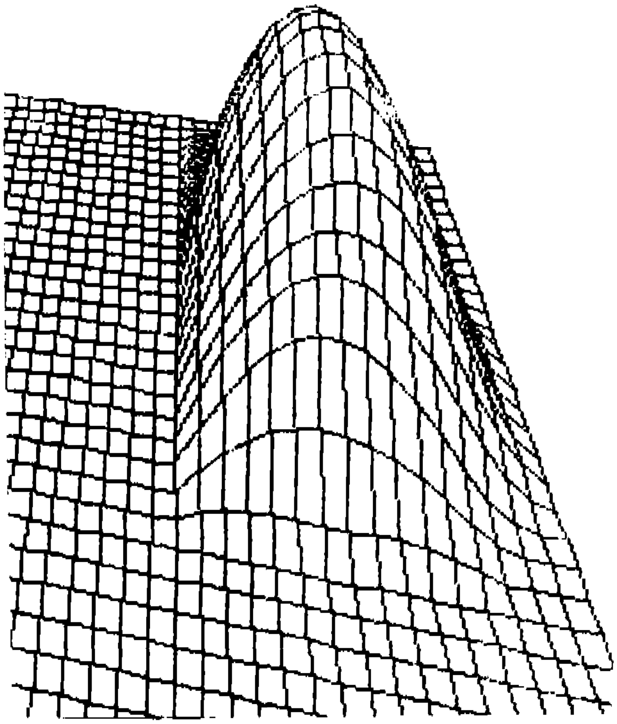


Fig. 3a

$x_{i-1}=0,\ x_i=1,\ x_{i+1}=x_{i+2}=4$

$y_{j-1}=0,\ y_j=3,\ y_{j+1}=5,\ y_{j+2}=6$

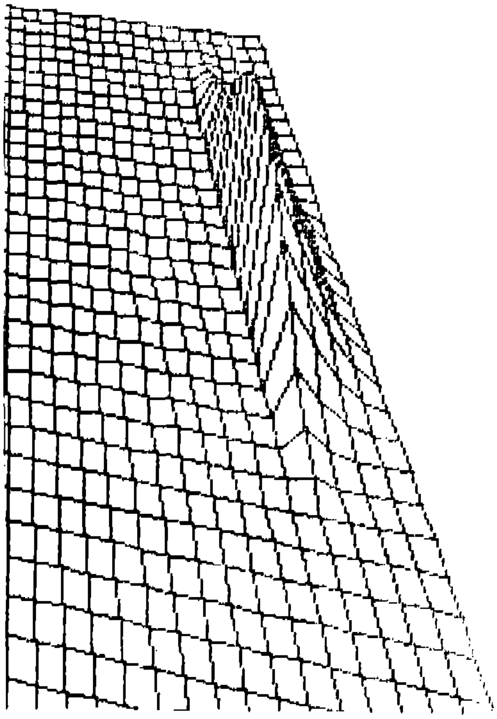


Fig. 3b

$x_{i-1}=0,\ x_i=x_{i+1}=x_{i+2}=1$

$y_{j-1}=0,\ y_j=3,\ y_{j+1}=5,\ y_{j+2}=6$

The provision of coalescence of the grid lines makes B_{ij} more flexible for application. It is interesting to point out that, different from the univariate setting, the B_{ij}'s are linearly dependent. In fact, we have the relationship

$$\sum_{j=-1}^{n+1}\sum_{i=-1}^{m+1}(x_{i+1}-x_i)(y_{j+1}-y_j)B_{ij}(x,y)=0$$

for all (x,y) in [a,b] × [c,d]. This dependence, however, does not diminish the utility of B_{ij} in various approximation schemes. We finally remark in this section that the center crisscross in the support of B_{ij} is in general active, and becomes inactive when the partition becomes regular. For regular grid partition, B_{ij} becomes a box-spline as discussed in [1] and this box-spline was first constructed in [9].

4. APPLICATION TO IMAGE RECONSTRUCTION. During the presentation at this "First Army Conference on Applied Mathematics and computing", we have shown C^1 images constructed from 64 × 64 and 128 × 128 discrete data of pictures of a girl, a portion of the surface of the moon, and a couple by simply using the surface

$$z = \sum_{j=-1}^{n+1} \sum_{i=-1}^{m+1} \frac{1}{4} (f(x_i,y_j) + f(x_i,y_{j+1}) + f(x_{i+1},y_j)$$

$$+ f(x_{i+1},y_{j+1}))B_{ij}(x,y)$$

where $x_{-2} = x_{-1} = x_0$, $x_{m+1} = x_{m+2} = x_{m+3}$, $y_{-2} = y_{-1} = y_0$, $y_{n+1} = y_{n+2} = y_{n+3}$, and the other x_i's and y_j's being equally spaced. This surface is compared to interpolating surfaces using bilinear, bicubic, and bicubic Hermite splines, and improves each of these tensor-product surfaces by at least 0.2 db, although it requires less computer time than the bicubic interpolating tensor-product surface. We expect that if variable grid lines $x = x_i$ and $y = y_j$ are used the images would greatly improve. More research and experiments are required in this direction.

REFERENCES

1. C. de Boor, Multivariate B-splines, this Proceedings.

2. C. de Boor and K. Höllig, Bivariate box splines and smooth pp functions on a three-direction mesh. MRC Report #2415, 1982.

3. C. K. Chui and R. H. Wang, On smooth multivariate spline spaces, Math. Comp. To appear.

4. C. K. Chui and R. H. Wang, Multivariate B-splines on triangulated rectangles, J. Math. Analysis and Appl. 92 (1983), 533-551.

5. C. K. Chui and R. H. Wang, Multivariate spline spaces. J. Math. Analysis and Appl. To appear.

6. C. K. Chui and R. H. Wang, Spaces of bivariate cubic and quartic splines on type-1 triangulations, J. Math. Analysis and Appl. To appear.

7. C. K. Chui and R. H. Wang, Bivariate B-splines on triangulated rectangles, in Approximation Theory IV, Ed. by C. K. Chui, L. L. Schumaker, and J. D. Ward, Academic Press, N.Y. 1983.

8. W. Dahmen and C. A. Micchelli, Recent progress in multivariate splines, in Approximation Theory IV, Ed. by C. K. Chui, L. L. Schumaker, and J. D. Ward, Academic Press, N.Y. 1983.

9. P. Zwart, Multivariate splines with nondegenerate partitions, SIAM Numer. Anal. 10 (1973), 665-673.

SPECTRAL METHODS FOR PARTIAL DIFFERENTIAL EQUATIONS

M. Yousuff Hussaini
Institute for Computer Applications in Science and Engineering
NASA Langley Research Center, Hampton, VA 23665

Craig L. Streett and Thomas A. Zang
NASA Langley Research Center, Hampton, VA 23665

ABSTRACT. Origins of spectral methods, especially their relation to the Method of Weighted Residuals, are surveyed. Basic Fourier, Chebyshev, and Legendre spectral concepts are reviewed, and demonstrated through application to simple model problems. Both collocation and tau methods are considered. These techniques are then applied to a number of difficult, nonlinear problems of hyperbolic, parabolic, elliptic, and mixed type. Fluid-dynamical applications are emphasized.

I. INTRODUCTION. Spectral methods may be viewed as an extreme development of the class of discretization schemes known by the generic name of the method of weighted residuals (MWR) [1]. The key elements of the MWR are the trial functions (also called the expansion or approximating functions) and the test functions (also known as weight functions). The trial functions are used as the basis functions for a truncated series expansion of the solution, which, when substituted into the differential equation, produces the residual. The test functions are used to enforce the minimization of the residual.

The choice of trial functions is what distinguishes the spectral methods from the finite element and finite difference methods. The trial functions for spectral methods are infinitely differentiable global functions. (Typically they are tensor products of the eigenfunctions of singular Sturm-Liouville problems.) In the case of finite element methods, the domain is divided into small elements, and a trial function is specified in each element. The trial functions are thus local in character, and well-suited for handling complex geometries. The finite difference trial functions are likewise local.

The choice of test function distinguishes between the Galerkin, collocation, and tau approaches. In the Galerkin approach, the test functions are the same as the trial functions, whereas in the collocation approach the test functions are translated Dirac delta functions. In other words, the Galerkin approach is equivalent to a least squares approximation, whereas the collocation approach requires the differential equation to be satisfied exactly at the collocation points. Spectral tau methods are close to Galerkin methods but they differ in the treatment of boundary conditions.

The collocation approach is the simplest of the MWR, and appears to have been first used by Slater [2] in his study of electronic energy bands in metals. A few years later, Barta [3] applied this method to the problem of the torsion of a square prism. Frazer, et al. [4] developed it as a general method for solving ordinary differential equations. They used a variety of trial functions and an arbitrary distribution of collocation points. The work of Lanczos [5] established for the first time that a proper choice of trial functions and distribution of collocation points is crucial to the accuracy of the solution. Perhaps he should be credited with laying down the foundation of the orthogonal collocation method. This method was revived by Clenshaw [6], Clenshaw and Norton [7], and Wright [8]. These studies involved application of Chebyshev polynomial expansions to initial value problems. Villadsen and Stewart [9] developed this method for boundary value problems.

The earliest investigations of the spectral collocation method to partial differential equations were those of Kreiss and Oliger [10] (who called it the Fourier method) and Orszag [11] (who termed it pseudospectral). This approach is especially attractive because of the ease with which it can be applied to variable coefficient and even nonlinear problems. The essential details will be furnished below.

The Galerkin approach is perhaps the most esthetically pleasing of the MWR since the trial functions and the test functions are the same. Indeed, the first serious application of spectral methods to PDE's -- that of Silberman [12] for meteorological modelling -- used the Galerkin approach. However, spectral Galerkin methods only became practical for high resolution calculations of nonlinear problems after Orszag [13] and Eliasen, et al. [14] developed a transform method for evaluating convolution sums arising from quadratic nonlinearities. Even in this case spectral collocation methods retain a factor of 2 in speed. For more complicated nonlinear terms high resolution spectral Galerkin methods are still impractical.

The tau approach is the most difficult to rationalize within the context of the MWR. Lanczos [5] developed the spectral tau method as a modification of the Galerkin method for problems with non-periodic boundary conditions. Although it too, is difficult to apply to nonlinear problems, it has proven quite useful for constant coefficient problems or subproblems, e.g., for semi-implicit time-stepping algorithms.

The following discussion of spectral methods for PDE's will be organized around the three basic types of systems -- hyperbolic, parabolic, and elliptic -- with an additional section for a difficult, nonlinear problem of mixed type. Simple, one-dimensional, linear examples will be provided to illustrate the basic principles and details of the algorithms; two-dimensional, nonlinear examples drawn from fluid dynamical applications will also be furnished to demonstrate the power of the method. The focus will be on collocation methods, although some discussion of tau methods is provided.

II. HYPERBOLIC EQUATIONS.

Linear hyperbolic equations are perhaps the simplest setting for describing spectral collocation methods. Both Fourier and Chebyshev schemes have found wide application. This section will first present the fundamentals of both approaches and then illustrate them on a nonlinear fluid dynamics problem involving shock waves.

Basic Fourier Spectral Concepts.

The potential accuracy of spectral methods derives from their use of suitable high-order interpolation formulae for approximating derivatives. An elementary example is provided by the model problem

$$u_t + u_x = 0, \tag{1}$$

with periodic boundary conditions on $[0,2\pi]$ and the initial condition

$$u(x,0) = \sin(\pi \cos x). \tag{2}$$

The exact solution

$$u(x,t) = \sin[\pi \cos(x-t)] \tag{3}$$

has the Fourier expansion

$$u(x,t) = \sum_{k=-\infty}^{\infty} \bar{u}_k(t)\, e^{ikx}, \tag{4}$$

where the Fourier coefficients

$$\bar{u}_k(t) = \sin\left(\frac{k\pi}{2}\right) J_k(\pi)\, e^{-ikt} \tag{5}$$

and $J_k(t)$ is the Bessel function of order k. The asymptotic properties of the Bessel functions imply that

$$k^p\, \bar{u}_k(t) \to 0 \qquad \text{as} \quad k \to \infty \tag{6}$$

for all positive integers p. As a result, the truncated Fourier series

$$u_N(x,t) = \sum_{k=-N/2+1}^{N/2-1} \bar{u}_k(t)\, e^{ikx} \tag{7}$$

converges faster than any finite power of $1/N$. This property is often referred to as exponential convergence. A straightforward integration-by-parts argument [15] may be used to show that it applies to any periodic and infinitely differentiable solution.

The standard collocation points are

$$x_j = \frac{2\pi j}{N} \qquad j=0,1,\cdots,N-1. \tag{8}$$

Let u_j denote the approximation to $u(x_j)$, where the time dependence has been suppressed. Then the spatial discretization of Eq. (1) is

$$\frac{\partial u_j}{\partial t} = \frac{\partial \tilde{u}}{\partial x}\bigg|_j , \tag{9}$$

where the right-hand-side is determined as follows. First, compute the discrete Fourier coefficients

$$\hat{u}_k = \frac{1}{N} \sum_{j=0}^{N-1} u_j \, e^{-ikx_j}, \qquad k = -\frac{N}{2}, -\frac{N}{2}+1, \cdots, \frac{N}{2}-1. \tag{10}$$

Then the interpolating function

$$\tilde{u}(x) = \sum_{k=-N/2}^{N/2-1} \hat{u}_k \, e^{ikx} \tag{11}$$

can be differentiated analytically to obtain

$$\frac{\partial \tilde{u}}{\partial x}\bigg|_j = \sum_{k=-N/2+1}^{N/2-1} ik \, \hat{u}_k \, e^{ikx_j}. \tag{12}$$

(The term involving $k = -N/2$ makes a purely imaginary contribution to the sum and hence has been dropped.) Note that each derivative approximation uses all available information about the function values. The sums in Eqs. (10) and (12) can be obtained in $O(N \ell n\ N)$ operations via the Fast Fourier Transform (FFT).

An illustration of the superior accuracy available from the spectral method for this problem is provided in Table I. Shown there are the maximum errors at $t = 1$ for the truncated series and for the spectral collocation method as well as for second-order and fourth-order finite difference methods. The time discretization was the classical fourth-order Runge-Kutta method. In all cases the time-step was chosen so small that the temporal discretization error was negligible. Because the solution is infinitely smooth, the convergence of the spectral method on this problem is more rapid than any finite power of $1/N$. (The error for the $N = 64$ spectral result is so small that it is swamped by the round-off error of these single precision CDC Cyber 175 calculations.) In most practical applications the benefit of the spectral method is not the extraordinary accuracy available for large N but rather the small size of N necessary for a moderately accurate solution.

Table I. Maximum Error for a 1-D Periodic Problem

N	Truncated Series	Fourier Spectral	2nd-Order Finite Difference	4th-Order Finite Difference
8	9.87 (-2)	1.62 (-1)	1.11 (0)	9.62 (-1)
16	2.55 (-4)	4.97 (-4)	6.13 (-1)	2.36 (-1)
32	1.05 (-11)	1.03 (-11)	1.99 (-1)	2.67 (-2)
64	6.22 (-13)	9.55 (-12)	5.42 (-2)	1.85 (-3)
128			1.37 (-2)	1.18 (-4)

Basic Chebyshev Spectral Concepts. Spectral methods for non-periodic problems can also exhibit exponential convergence. A simple example is again provided by Eq. (1) but now on the interval [-1,1] with initial condition u(x,0) and boundary condition u(-1,t). Since this is not a periodic problem, a spectral method based upon Fourier series in x would exhibit extremely slow convergence. However, rapid convergence as well as efficient algorithms can be attained for spectral methods based upon Chebyshev polynomials. These are defined on [-1,1] by

$$\tau_n(x) = \cos\,(n \cos^{-1} x). \tag{13}$$

The function

$$u(x,t) = \sin\, \alpha\pi(x-t) \tag{14}$$

is one solution to Eq. (1). It has the Chebyshev expansion

$$u(x,t) = \sum_{n=0}^{\infty} \bar{u}_n(t)\, \tau_n(x), \tag{15}$$

where

$$\bar{u}_n(t) = \frac{2}{c_n} \sin\left(\frac{n\pi}{2} - \alpha\pi t\right) J_n(\alpha\pi) \tag{16}$$

with

$$c_n = \begin{cases} 2 & n = 0 \\ 1 & n \geqslant 1 \end{cases} . \tag{17}$$

The truncated series

$$u_N(x,t) = \sum_{n=0}^{N} \bar{u}_n(t)\, \tau_n(x) \tag{18}$$

converges at an exponential rate. Note that this result holds whether or not α is an integer. In contrast, the Fourier coefficients of $u(x,t)$ are

$$\bar{u}_k(t) = \frac{i}{2\pi} e^{i\alpha\pi t} \frac{\sin \pi(\alpha+k)}{\alpha+k} - \frac{i}{2\pi} e^{-i\alpha\pi t} \frac{\sin \pi(\alpha-k)}{\alpha-k} . \tag{19}$$

For non-integer α these decay extremely slowly.

The change of variables

$$x = \cos \theta, \tag{20}$$

the definition

$$v(\theta,t) = u(\cos \theta,t), \tag{21}$$

and Eq. (13) reduce Eq. (15) to

$$v(\theta,t) = \sum_{n=0}^{\infty} \bar{u}_n(t) \cos n\theta. \tag{22}$$

Thus, the Chebyshev coefficients of $u(x,t)$ are precisely the Fourier coefficients of $v(\theta,t)$. This new function is automatically periodic. If $u(x,t)$ is infinitely differentiable (in x), then $v(\theta,t)$ will be infinitely differentiable (in θ). Hence, straightforward integration-by-parts arguments lead to the conclusion that the Chebyshev coefficients of an infinitely differentiable function will decay exponentially fast. Note that this holds regardless of the boundary conditions.

A Chebyshev spectral method makes use of the interpolating function

$$\tilde{u}(x) = \sum_{n=0}^{N} \hat{u}_n \tau_n(x). \tag{23}$$

The standard collocation points are

$$x_j = \cos \frac{\pi j}{N} \qquad j = 0,1,\cdots,N. \tag{24}$$

Thus,

$$u_j = \sum_{n=0}^{N} \hat{u}_n \cos \frac{n\pi j}{N} , \tag{25}$$

where u_j is the approximation to $u(x_j)$. The inverse relation is

$$\hat{u}_n = \frac{2}{N\bar{c}_n} \sum_{j=0}^{N} \bar{c}_j^{-1} u_j \cos \frac{n\pi j}{N}, \qquad n = 0,1,\cdots,N \tag{26}$$

where

$$\bar{c}_j = \begin{cases} 2 & j = 0 \quad \text{or} \quad N \\ 1 & 1 \leqslant j \leqslant N-1 \end{cases} . \tag{27}$$

The analytic derivative of this function is

$$\frac{\partial \tilde{u}}{\partial x} = \sum_{n=0}^{N} \hat{u}_n^{(1)} \tau_n(x), \tag{28}$$

where

$$\begin{aligned} \hat{u}^{(1)}_{N+1} &= 0 \\ \hat{u}^{(1)}_{N} &= 0 \\ \bar{c}_n \hat{u}^{(1)}_n &= \hat{u}^{(1)}_{n+2} + 2(n+1)\hat{u}_{n+1}, \qquad n = N-1,N-2,\cdots,0. \end{aligned} \tag{29}$$

(See [15] for the derivation of this recursion relation.) The Chebyshev spectral derivatives at the collocation points are

$$\left.\frac{\partial \tilde{u}}{\partial x}\right|_j = \sum_{n=0}^{N} \hat{u}_n^{(1)} \cos \frac{\pi jn}{N} . \tag{30}$$

Special versions of the FFT may be used for evaluating the sums in Eqs. (26) and (30). The total cost for a Chebyshev spectral derivative is thus $O(N \ell n N)$.

The time-stepping scheme for Eq. (1) must use the boundary conditions to update u_N (at $x = -1$) and the approximate derivatives from Eq. (30) to update u_j for $j=0,1,\cdots,N-1$. Note that no special formula is required for the derivative at $j = 0$ (or $x = +1$).

Results pertaining to $\alpha = 2.5$ at $t = 1$ for a truncated Chebyshev series, a Chebyshev spectral method, a Fourier spectral method, and a second-order finite difference method are given in Table II. For this non-periodic problem Fourier spectral methods are quite inappropriate, but the Chebyshev spectral method is far superior to the finite difference method.

The Chebyshev collocation points are the extreme points of $\tau_N(x)$. Note that they are not evenly distributed in x, but rather are clustered near the endpoints. The smallest mesh size scales as $1/N^2$. While this distribution contributes to the quality of the Chebyshev approximation and permits the use of the FFT in evaluating the series, it also places a severe time-step limitation on explicit methods for evolution equations.

Table II. Maximum Error for a 1-D Dirichlet Problem

N	Truncated Series	Chebyshev Spectral	Fourier Spectral	Finite Difference
4	1.24 (0)	1.49 (0)	1.85 (0)	1.64 (0)
8	1.25 (-1)	6.92 (-1)	1.92 (0)	1.73 (0)
16	7.03 (-6)	1.50 (-4)	2.27 (0)	1.23 (0)
32	1.62 (-13)	3.45 (-11)	2.28 (0)	3.34 (-1)
64	1.79 (-13)	9.55 (-11)	2.27 (0)	8.44 (-2)

Application to Two-dimensional, Supersonic Flow. Spectral methods have recently been applied successfully to the nonlinear hyperbolic system of equations which describes a two-dimensional inviscid gas [16,17]. The most serious complication over the simple model problems discussed above occurs when shock waves are present. If the shock occurs in the interior of the domain, then the truncated series for the discontinuous flow variables converges very slowly. Elaborate filtering strategies appear necessary to extract useful information from a calculation of such a situation [17,18]. This difficulty disappears, however, when the shock occurs at the boundary of the domain, as in shock-fitting as opposed to shock-capturing calculations.

A schematic of the type of spectral shock-fitted calculations described below is illustrated in Fig. 1. At time $t = 0$ an infinite, normal shock at $x = 0$ separates a rapidly moving, uniform fluid on the left from the fluid on the right which is in a quiescent state except for some specified fluctuation. The initial conditions are chosen so that in the absence of any fluctuation the shock moves uniformly in the positive x-direction with a Mach number (relative to the fluid on the right) denoted by M_s. In the presence of fluctuations the shock front will develop ripples. The shape of the shock is described by the function $x_s(y,t)$. The numerical calculations are used to determine the state of the fluid in the region between the shock front and some suitable left boundary $x_L(t)$ and also to determine the motion and shape of the shock front itself.

Figure 1 is taken from a shock/turbulence calculation [19] in which the downstream fluctuation is a plane vorticity wave that is periodic in y with period y_ℓ. Because of the initial value nature of the calculation, the fluid motion behind the shock is not periodic in x, as Fig. 1 makes abundantly clear. The interesting physical domain is given by

$$\begin{gathered} x_L(t) \leqslant x \leqslant x_s(y,t) \\ 0 \leqslant y \leqslant y_\ell \\ t \geqslant 0. \end{gathered} \tag{31}$$

The change of variables

$$\begin{aligned} X &= \frac{x - x_L(t)}{x_s(y,t) - x_L(t)} \\ Y &= y/y_\ell \\ T &= t \end{aligned} \tag{32}$$

produces the computational domain

$$\begin{cases} 0 \leqslant X \leqslant 1 \\ 0 \leqslant Y \leqslant 1 \\ T \geqslant 0. \end{cases} \tag{33}$$

The fluid motion is modeled by the two-dimensional Euler equations. In terms of the computational coordinates these are

$$Q_T + B\,Q_X + C\,Q_Y = 0, \tag{34}$$

where $Q = (P,u,v,S)^T$,

$$B = \begin{bmatrix} U & \gamma X_x & \gamma X_y & 0 \\ \frac{a^2}{\gamma} X_x & U & 0 & 0 \\ \frac{a^2}{\gamma} X_y & 0 & U & 0 \\ 0 & 0 & 0 & U \end{bmatrix} \tag{35}$$

and

$$C = \begin{bmatrix} V & \gamma Y_x & \gamma Y_y & 0 \\ \frac{a^2}{\gamma} Y_x & V & 0 & 0 \\ \frac{a^2}{\gamma} Y_y & 0 & V & 0 \\ 0 & 0 & 0 & V \end{bmatrix} . \tag{36}$$

The contravariant velocity components are given by

$$U = X_t + uX_x + vX_y$$

and

$$V = Y_t + uY_x + vY_y . \tag{37}$$

A subscript denotes partial differentiation with respect to the indicated variable. P, a, and S are all normalized by reference conditions at downstream infinity; u and v are velocity components in the x and y directions, both scaled by the characteristic velocity defined by the square root of the pressure-density ratio at downstream infinity. A value $\gamma = 1.4$ has been used.

Let n denote the time level and Δt the time increment. The time discretization of Eq. (34) is

$$\tilde{Q} = [1 - \Delta t L^n] Q^n \tag{38}$$

$$Q^{n+1} = \frac{1}{2} [Q^n + (1 - \Delta t \tilde{L}) \tilde{Q}] , \tag{39}$$

where L denotes the spatial discretization of $B\partial_X + C\partial_Y$. The solution Q has the Chebyshev - Fourier series expansion

$$Q(X,Y,T) = \sum_{p=0}^{M} \sum_{q=-N/2}^{N/2-1} Q_{pq}(T) \, \tau_p(\xi) e^{2\pi i q Y} , \tag{40}$$

where $\xi = 2X-1$. The derivatives Q_X and Q_Y are approximated by

$$Q_X = 2 \sum_{p=0}^{M} \sum_{q=-N/2}^{N/2-1} Q_{pq}^{(1,0)}(T) \tau_p(\xi) e^{2\pi i q Y} , \tag{41}$$

$$Q_Y = 2\pi \sum_{p=0}^{M} \sum_{q=-N/2}^{N/2-1} Q_{pq}^{(0,1)}(T) \tau_p(\xi) e^{2\pi i q Y} , \tag{42}$$

where $Q_{pq}^{(1,0)}$ is computed from Q_{pq} in a manner analogous to Eq. (29), and

$$Q_{pq}^{(0,1)} = i \, q \, Q_{pq} . \tag{43}$$

As a general rule the correct numerical boundary conditions for a spectral method are the same as the correct analytical boundary conditions. The global nature of the approximation avoids the need for special differentiation formulae at boundaries. At the same time spectral methods are quite unforgiving of incorrect boundary conditions. The inherent dissipation of these methods is so low that boundary errors quickly contaminate the entire solution. In many fluid dynamical applications the computational region must be terminated at some finite, artificial boundary. The difficulty at "artificial" boundaries is that analytically correct, fully nonlinear boundary conditions for systems are seldom known. One example of a workable artificial boundary condition for the Euler equations is given in Ref. [20].

The most critical part of the calculation is the treatment of the shock front. The shock-fitting approach used here is desirable because it avoids the severe post-shock oscillations that plague shock-capturing methods. The time derivative of the Rankine-Hugoniot relations provides an equation for the shock acceleration. This equation is integrated to update the shock position (see [20] for details). This method is a generalization of the finite difference method developed by Pao and Salas [21] for their study of the shock/vortex interaction.

The nonlinear interaction of plane waves with shocks was examined at length in [19]. The numerical method used there was similar to the one described above but employed second-order finite differences in place of the present Chebyshev-Fourier spectral discretization. Detailed comparisons were made in [19] with the predictions of linear theory [22]. The linear results turned out to be surprisingly robust, remaining valid at very low (but still supersonic) Mach numbers and at very high incident wave amplitudes. The only substantial disagreement occurred for incident waves whose wave fronts were nearly perpendicular to the shock front. This type of shock-turbulence interaction is a useful test of the spectral technique because the method can be calibrated in the regions for which linear theory has been shown to be valid.

The most reliable numerical results can be obtained for the acoustic responses to acoustic waves. Unlike the vorticity responses, these require no differentiation of the flow variables, thus eliminating one extra source of error. Moreover, the acoustic reponse stretches much further behind the shock than the vorticity response, thus providing greater statistical reliability. Vorticity response results are reported in [23]. The incident pressure wave is taken to be

$$p_1' = A_1' \, e^{i(\underline{k}_1 \cdot \underline{x} - \omega_1 t)} \tag{44}$$

where $\underline{k}_1 = (k_{1,x}, k_{1,y})$, $\omega = M_s a_1 \, k_{1,x} + a_1 \, k_1$ and A_1' is the amplitude. In terms of the incidence angle θ_1, $\underline{k}_1 = (k_1 \cos \theta_1, k_1 \sin \theta_1)$.

The linearized transmitted acoustic wave can be expressed in the same manner with all subscripts changed from 1 to 2. The amplification coefficient for the transmitted acoustic wave is then the ratio

$$A_2^{\prime}/A_1^{\prime} \ . \tag{45}$$

Figure 2 indicates the transmission coefficient extracted from the computation. At each fixed value of X we perform a Fourier analysis in Y of the pressure. The Fourier coefficient for $q = 1$ provides the amplitude $A_2^{\prime}$. In order to reduce the transients that would accompany an abrupt start of the calculation at full wave amplitude, an extra factor of $s(t)$ is inserted into Eq. (44), where

$$s(t) = \begin{cases} 3(t/t_s)^2 - 2(t/t_s)^3 & 0 \leqslant t \leqslant t_s \\ 1 & t \geqslant t_s \end{cases} \ . \tag{46}$$

The start-up time t_s is some multiple (typically $\frac{1}{2}$) of the time it takes the shock to encounter one full wavelength (in the x-direction) of the incident wave. The ratio $A_2^{\prime}/A_1^{\prime}$ is plotted in Fig. 2 as a function of the mean value of the physical coordinate x corresponding to X. The start-up time for this Mach 3 case is $t_s = 0.56$. The average of the x-dependent responses between the start-up interval and the shock produces the computed transmission coefficient. The standard deviation of the individual responses serves as an error estimate.

The dependence upon incidence angle of the acoustic transmission coefficient for $A_1^{\prime} = 0.001$ and $M_s = 3$ waves is displayed in Fig. 3. As is discussed in [19], linear theory is quite reliable at angles below, say, 45°. Figure 3 contains results from both spectral and finite difference calculations. The finite difference results were obtained with the same second-order MacCormack's method that was described in [19] except that periodic boundary conditions (rather than stretching) were employed in the y-direction. The finite difference grid was 64 × 16 and these calculations used a CFL number of 0.70. The spectral grid was 32 × 8,and the CFL number was 0.50. Figure 3 shows that both methods produce the same results. A head-to-head comparison of both methods for the $\theta_1 = 10^{\circ}$ case is provided in Table III. The "exact" value is taken from linear theory [22]. Since the amplitude of the incident acoustic wave is so small, it should come as no surprise that four points in the y-direction suffice for the spectral calculation. Note that the standard deviations are substantially smaller for the spectral method. These results suggest that the spectral method requires only half as many grid points in each coordinate direction.

Table III. Grid Dependence of Acoustic Transmission Coefficient

Grid	Finite Difference	Chebyshev-Fourier Spectral
16 × 4	6.403 ± 2.652	7.257 ± 0.587
16 × 8	6.427 ± 2.626	7.257 ± 0.587
32 × 4	7.105 ± 0.453	7.158 ± 0.022
32 × 8	7.134 ± 0.471	7.158 ± 0.022
32 × 16	7.139 ± 0.497	7.158 ± 0.022
64 × 16	7.163 ± 0.078	7.157 ± 0.017
128 × 16	7.152 ± 0.022	
"exact"	7.156	7.156

III. PARABOLIC EQUATIONS. The nonlinear, parabolic system formed by the incompressible, Navier-Stokes equations was the focus of much of the early development and application of spectral methods to large-scale fluid dynamical problems. Fourier spectral methods have been the obvious choice for the simulation of homogeneous, isotropic turbulence [24]. For shear flows, however, non-periodic boundary conditions are required. So far, Chebyshev spectral methods have been favored for these applications [25,26,27]. Nevertheless, Legendre spectral methods are a viable alternative and of late they have been attracting some attention. This section will present a discussion of the implementation of Legendre spectral methods and will then compare them with Chebyshev spectral methods for the one-dimensional heat equation. This section will close with a description of a promising semi-implicit time-stepping scheme for the Navier-Stokes equations.

Basic Legendre Spectral Concepts. A Legendre spectral method on [-1,1] makes uses of the interpolating function

$$\tilde{u}(x) = \sum_{n=0}^{N} \hat{u}_n P_n(x), \tag{47}$$

where $P_n(x)$ is the Legendre polynomial of degree n. Closed form expressions for these polynomials are well-known, albeit clumsy. The computationally preferred way to evaluate the polynomials is through the recursion relation

$$P_0(x) = 1$$

$$P_1(x) = x$$

and for $n \geqslant 2$

$$n\, P_n(x) = (2n-1)xP_{n-1}(x) - (n-1)\, P_{n-2}(x). \tag{48}$$

Unlike the case with Fourier and Chebyshev collocation methods, there is no tidy expression for the Legendre collocation points. Appeal must be made to the theory of numerical quadrature [28]. The presence of boundary conditions at both endpoints makes it desirable to include -1 and $+1$ in the set of collocation points. Subject to this constraint, the most accurate quadrature formula for the discrete Legendre coefficients is the Gauss-Lobatto rule

$$\hat{u}_n = \hat{c}_n \sum_{j=0}^{N} w_j \, P_n(x_j) \, u_j, \qquad n = 0,1,\cdots,N \qquad (49)$$

where $x_0 = +1$, $x_N = -1$ and x_j for $1 \leqslant j \leqslant N-1$ are the roots of $P_N'(x)$. The weights are

$$w_j = \frac{1}{N(N+1) \, P_N^2(x_j)} , \qquad (50)$$

and

$$\hat{c}_n = \begin{cases} 2n+1 & n = 0,1,\cdots,N-1 \\ N & n = N \end{cases} . \qquad (51)$$

The interior collocation points must be determined numerically. This quadrature rule yields the exact Legendre coefficients if $u(x)$ is any polynomial of degree less than N. Its inverse relation is

$$u_j = \sum_{n=0}^{N} \hat{u}_n \, P_n(x_j). \qquad (52)$$

The analytic derivative of the interpolating function in Eq. (47) is

$$\frac{\partial \tilde{u}}{\partial x} = \sum_{n=0}^{N} \hat{u}_n^{(1)} \, P_n(x), \qquad (53)$$

where

$$\begin{aligned} &\hat{u}_{N+1}^{(1)} = 0 \\ &\hat{u}_N^{(1)} = 0 && \qquad (54) \\ &\frac{1}{2n+1} \hat{u}_n^{(1)} = \frac{1}{2n+5} \hat{u}_{n+2}^{(1)} + \hat{u}_{n+1} && \qquad n = N-1,N-2,\cdots,0. \end{aligned}$$

Since fast transform techniques are not available for the Legendre basis functions, there is no particular advantage to computing $\partial\tilde{u}/\partial x|_j$ by applying Eqs. (49), (54) and (53) rather than by following Eq. (49) with

$$\frac{\partial\tilde{u}}{\partial x}\Big|_j = \sum_{n=0}^{N} \hat{u}_n \, P_n'(x_j). \tag{55}$$

In fact, for a collocation method it is faster still to perform this entire process by a single matrix-vector multiplication. For that matter the Chebyshev collocation differentiation operator may also be represented by a matrix. Timing studies [29] on the CDC Cyber 175 have indicated that even for N = 16, the Chebyshev matrix-multiply differentiation procedure is substantially faster than one based on assembly language fast transforms. Moreover, the matrix-multiply procedure does not suffer the sort of speed degradation that afflicts the transform procedure whenever N is not an integral power of 2.

The heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{56}$$

is the natural parabolic linear model problem. The spatial domain is [-1,1], the initial condition is

$$u(x,0) = \sin \pi x \tag{57}$$

and the boundary conditions are

$$\begin{aligned} u(-1,t) &= 0 \\ u(+1,t) &= 0. \end{aligned} \tag{58}$$

The exact solution is then

$$u(x,0) = e^{-\pi^2 t} \sin \pi x. \tag{59}$$

The time differencing is again the classical fourth-order Runge-Kutta scheme.

In addition to spectral collocation and series truncation solutions, we will also present spectral tau results. Let $\bar{u}_n(t)$ for $n=0,1,\cdots,N$ denote the Legendre coefficients of the tau approximation to $u(x,t)$. The semi-discrete tau equations are

$$\frac{d\bar{u}_n}{dt} = \bar{u}_n^{(2)}, \qquad n = 0,1,\cdots,N-2 \tag{60}$$

with

$$\sum_{\substack{n=0 \\ n \text{ even}}}^{N} \bar{u}_n = 0 \tag{61}$$

$$\sum_{\substack{n=1 \\ n \text{ odd}}}^{N} \bar{u}_n = 0.$$

The Legendre coefficients of the approximation to the second spatial derivative $\bar{u}_n^{(2)}(t)$ can be obtained from $\bar{u}_n(t)$ by two applications of the recursion relation in Eq. (54). In this tau approximation the dynamical equations for the two highest-order coefficients are dropped in favor of the equations for the boundary conditions. Equation (61) follows from the property

$$P_n(\pm 1) = (\pm 1)^n. \tag{62}$$

Since the Chebyshev polynomials also satisfy Eq. (62), the Chebyshev tau equations for Eq. (56) are the same as Eqs. (60) and (61). Of course, Eq. (29) is invoked for the derivative coefficients instead of Eq. (54).

The results at $t = 1$ are given in Tables IV and V. The maximum errors shown there have been boosted up by the factor e^{π^2} so that they represent relative errors. On the whole the collocation results are the best. Moreover, except for the truncated series results, the Legendre approximations are superior to the Chebyshev ones. Lanczos [30] has discussed some circumstances under which Legendre approximations are superior to Chebyshev ones. It goes almost without saying that finite difference results are far inferior to any of these spectral approximations.

Table IV. Maximum Error for Legendre Approximations to the Heat Equation

N	Truncated Series	Tau	Collocation
8	6.65 (-4)	6.85 (-4)	2.40 (-5)
10	1.72 (-5)	1.07 (-5)	1.50 (-7)
12	3.06 (-7)	1.54 (-7)	1.38 (-9)
14	3.50 (-9)	1.86 (-9)	4.81 (-10)
16	3.88 (-11)	1.15 (-10)	9.98 (-11)

Table V. Maximum Error for Chebyshev Approximations to the Heat Equation

N	Truncated Series	Tau	Collocation
8	2.44 (-4)	1.61 (-3)	4.58 (-4)
10	5.76 (-6)	2.12 (-5)	8.25 (-6)
12	9.42 (-8)	3.19 (-7)	1.01 (-7)
14	1.14 (-9)	3.35 (-9)	1.10 (-9)
16	1.05 (-11)	8.39 (-11)	2.09 (-11)

The time-step restriction for explicit Legendre or Chebyshev methods for the heat equation is very severe, scaling as $1/N^4$. This can pose quite a barrier to large-scale calculations for which a relative accuracy of 0.1% or so will suffice. Fortunately, many large-scale calculations can be split into one-dimensional, inhomogeneous counterparts of Eq. (56) and efficient implicit schemes are available for this linear, constant coefficient equation. They rely on reducing the Legendre (or Chebyshev) tau equations to a system which is nearly tridiagonal. The Legendre tau equations for a Crank-Nicolson temporal discretization of Eq. (56) are

$$\frac{\lambda}{(2n-1)(2n-3)}\,\bar{u}_{n-2} + \left[1 - \frac{2\lambda e_{n+2}}{(2n-1)(2n+3)}\right]\bar{u}_n + \frac{\lambda e_{n+4}}{(2n+3)(2n+5)}\,\bar{u}_{n+2}$$

$$= \frac{1}{(2n-1)(2n-3)}\,\bar{f}_{n-2} - \frac{2e_{n+2}}{(2n-1)(2n+3)}\,\bar{f}_n + \frac{e_{n+4}}{(2n+3)(2n+5)}\,\bar{f}_{n+2}$$

$$n = 2,3,\cdots,N, \qquad (63)$$

where $\lambda = -\Delta t/2$ with Δt the time-step, the coefficients $\bar{u}_n$ on the left-hand side are at $t + \Delta t$,

$$\bar{f}_n = \bar{u}_n(t) + \frac{1}{2}\,\Delta t\,\bar{u}_n^{(2)}(t), \qquad (64)$$

and

$$e_n = \begin{cases} 1 & 0 \leqslant n \leqslant N \\ 0 & n > N \end{cases} \quad . \qquad (65)$$

Equation (63) for even n plus the first of Eqs. (61) from a linear system which is tridiagonal except for the boundary condition equation. This is cheap to invert. The odd coefficients display a similar structure. The Chebyshev tau version of Eq. (63) is available in [15] and [31].

Application to Channel Flow. Several three-dimensional Navier-Stokes algorithms have been developed which incorporate the quasi-tridiagonal structure of the Chebyshev tau equations for the second derivative in semi-implicit schemes which treat the constant coefficient diffusion term implicitly [25-27]. In practice this device has permitted time-steps several orders of magnitude larger than the explicit diffusion limit. Unfortunately, the quasi-tridiagonal structure is lost even for a linear, variable viscosity coefficient. An effective iterative scheme for this more general case has recently been proposed by Malik, et al. [32]. This approach will be described here in its two-dimensional setting.

The rotation form equations for two-dimensional channel flow are

$$\begin{aligned} u_t - v(v_x - u_y) + P_x &= (\mu u_x)_x + (\mu u_y)_y \\ v_t + u(v_x - u_y) + P_y &= (\mu v_x)_x + (\mu v_y)_y \\ u_x + v_y &= 0, \end{aligned} \tag{66}$$

with periodic boundary conditions in x and no-slip boundary conditions at $y = \pm 1$. The variable P denotes the total pressure. The viscosity μ is presumed to depend upon y.

A useful discretization employs Fourier series in x and Chebyshev series in y. The pressure gradient term and the incompressibility constraint are best handled implicitly. So, too, are the vertical diffusion terms because of the fine mesh-spacing near the wall. The variable viscosity prevents the standard Poisson equation for the pressure from decoupling from the velocities in the diffusion term. The algorithm described in [26] appears to be a good starting point. A Crank-Nicolson approach is used for the implicit terms and Adams-Bashforth for the remainder. After a Fourier transform in x, the equations for each wavenumber k have the following implicit structure

$$\begin{aligned} \hat{u} - \tfrac{1}{2}\Delta t(\mu\hat{u}_y)_y + \tfrac{1}{2}\Delta t ik\hat{P} &= \cdots \\ \hat{v} - \tfrac{1}{2}\Delta t(\mu\hat{v}_y)_y + \tfrac{1}{2}\Delta t\hat{P}_y &= \cdots \\ ik\hat{u} + \hat{v}_y &= 0. \end{aligned} \tag{67}$$

Fourier transformed variables are denoted by hats, the subscript y denotes a Chebyshev spectral derivative, and Δt is the time increment.

The algorithm in [26] was devised for constant viscosity, in which case the Eqs. (67) can be reduced to essentially a block-tridiagonal form. This cannot be done in the present, more general situation. We advocate solving these equations iteratively after applying a finite difference pre-conditioning.

The interesting physical problems have high Reynolds number, i.e., low viscosity. Thus the first derivative terms in Eqs. (67) predominate. The effective pre-conditioning of them is crucial. Four possibilities have been considered. The eigenvalues of pre-conditioned iterations for the model scalar problem $u_x = f$ with periodic boundary conditions on $[0,2\pi]$ are given for each possibility in Table VI. The term $\alpha\Delta x$ is the product of a wavenumber α and the grid spacing Δx. It falls in the range $0 \leq |\alpha\Delta x| \leq \pi$. For the staggered grid case the discrete Eqs. (67) are modified so that the velocities and the momentum equations are defined at the cell faces $y_j = \cos(\pi j/N)$, $j=0,1,\cdots,N$, whereas the pressure and the continuity equation are defined at the cell centers $y_{j-1/2} = \cos(\pi(j-1/2)/N), j=1,\cdots,N$. Fast cosine transforms enable interpolation between cell faces and cell centers to be implemented efficiently. The staggered grid for the Navier-Stokes equations has the advantage that no artificial boundary condition is required for the pressure at the walls.

Table VI. Pre-conditioned Eigenvalues for a One-dimensional First Derivative Model Problem

PRE-CONDITIONING	EIGENVALUES
Central Differences	$\dfrac{\alpha\Delta x}{\sin(\alpha\Delta x)}$
One-sided Differences	$e^{-i(\alpha\Delta x/2)} \dfrac{\alpha\Delta x/2}{\sin((\alpha\Delta x)/2))}$
High Mode Cut-off	$\begin{cases} \dfrac{\alpha\Delta x}{\sin(\alpha\Delta x)} & 0 \leq \lvert\alpha\Delta x\rvert \leq (2\pi/3) \\ 0 & (2\pi/3) < \lvert\alpha\Delta x\rvert \leq \pi \end{cases}$
Staggered Grid	$\dfrac{(\alpha\Delta x)/2}{\sin((\alpha\Delta x)/2)}$

The actual eigenvalues for pre-conditioned iterations of Eqs. (67) are displayed in Fig. 4. The model problem estimates the eigenvalue trends surprisingly well considering that it is just a scalar equation, has only first derivative terms and uses Fourier series rather than Chebyshev polynomials.

The preceding results indicate that the staggered grid leads to the most effective treatment of the first derivative terms. The condition number of the pre-conditioned system is reasonably small and no resolution is lost by a high mode cut-off. (Although it is possible to devise a high-mode cut-off which avoids the small eigenvalues shown in the figures, some of the spectral resolution is thereby lost.) A simple and effective iterative scheme for this system with its complex eigenvalues is a minimum residual method [32]. At a Reynolds number of 7500 each iteration reduces the residual by almost an order of magnitude.

Table VII, which is taken from [32], presents a comparison of the accuracy of the Chebyshev discretization in y. The two codes are otherwise identical. The initial condition consisted of Poiseuille flow plus a small amount of a linearly unstable eigenmode. The table compares the computed growth rate of this perturbation with the theoretical, linear result after 100 time-steps.

Table VII. Percent Error in Growth Rate

N	Finite Difference	Spectral
8	4470	3210
16	337	74.5
32	147	0.097
64	39.5	0.071
128	10.0	
256	2.4	

IV. ELLIPTIC EQUATIONS. Fruitful nonlinear applications of spectral methods developed the latest for equations of elliptic type. Unlike hyperbolic or parabolic equations, for which explicit schemes can often be tolerated, ellipic equations virtually require implicit iterative schemes in practical situations. It was only a few years ago that Morchoisne [33] and Orszag [34] proposed preconditioning the spectral collocation equations by finite difference operators. More recently still, effective spectral multigrid iterative methods have been developed [35,36] and applied to the nonlinear potential flow problem of fluid dynamics [29]. These developments will be described in this section.

Poisson's Equation. As usual the discussion will begin with a linear model problem, but this time in two spatial dimensions. That problem is the Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f \tag{68}$$

on the square $[-1,1] \times [-1,1]$ with homogeneous Dirichlet boundary conditions. The choice

$$f(x,y) = -2\pi^2 \sin \pi x \sin \pi y \tag{69}$$

corresponds to the analytical solution

$$u(x,y) = \sin \pi x \sin \pi y. \tag{70}$$

Both Chebyshev and Legendre spectral methods are appropriate for this problem. Direct solution schemes for the Chebyshev tau method have been described in [31]. The same schemes also work for the Legendre tau method with straightforward modifications. They are basically of an alternating direction implicit (ADI) nature and rely on the quasi-tridiagonal form of the constant coefficient, one-dimensional problem. Haidvogel and Zang [31] report comparisons of the Chebyshev tau method with finite difference methods on numerous problems. They discuss both computational efficiency and accuracy.

These direct solution schemes cannot feasibly be extended to spectral collocation methods because the collocation equations for the one-dimensional components cannot be represented by sparse matrices. However, an ADI iterative scheme based on finite difference preconditioning is an efficient method for obtaining an approximate solution. The description of this scheme in its general nonlinear setting begins by writing the spectral collocation equations as

$$M(U) = 0. \tag{71}$$

Define the Jacobian

$$J(U) = \frac{\partial M}{\partial U}(U). \tag{72}$$

In many cases the Jacobian can be split into the sum of two operators $J_x(U)$ and $J_y(U)$, each involving derivatives in only the one coordinate direction indicated by the subscript. The most straightforward ADI method is

$$[\alpha I - J_x(V)][\alpha I - J_y(V)]\Delta V = \alpha M(V), \tag{73}$$

with the approximate solution V updated by

$$V \leftarrow V + \omega \Delta V. \tag{74}$$

This is just the Douglas-Gunn version of ADI [37]. The term approximate factorization is commonly used for this type of scheme for the nonlinear potential flow problem [38]. This particular scheme is referred to as AF1. For second-order spatial discretizations the term $[\alpha I - J_x(V)]$ leads to a set of tridiagonal systems, one for each value of y. The second left-hand side factor produces another set of tridiagonal systems. For spectral discretizations, however, these systems are full; hence, Eq. (73) is still relatively expensive to invert. A compromise is to replace J_x and J_y with their second-order finite difference analogs, denoted by H_x and H_y, respectively:

$$\left[\alpha I - H_x(V)\right]\left[\alpha I - H_y(V)\right]\Delta V = \alpha M(V). \tag{75}$$

The spectral approximate factorization scheme consists of Eqs. (74) and (75). The choice of the iteration parameters is discussed in [29].

Table VIII. Maximum Error for Chebyshev Approximations to Poisson's Equation

N	Truncated Series	Tau	Collocation
8	2.88 (-4)	2.79 (-3)	1.17 (-4)
10	6.79 (-6)	5.26 (-5)	2.33 (-6)
12	1.09 (-7)	8.86 (-7)	3.12 (-8)
14	1.34 (-9)	1.09 (-8)	3.27 (-10)
16	1.19 (-11)	9.15 (-11)	2.73 (-12)

The results for the simple model problem are presented in Tables VIII and IX. The trends are the same as they were for the heat equation: the collocation method is more accurate than tau and Legendre polynomials are more accurate than Chebyshev. (Since it is not practical to design a spectral method for PDE's using truncated series, those results have been ignored in this comparison.)

Table IX. Maximum Error for Legendre Approximations to Poisson's Equation

N	Truncated Series	Tau	Collocation
8	6.04 (-4)	1.55 (-3)	1.77 (-5)
10	1.69 (-5)	3.40 (-5)	2.48 (-7)
12	3.05 (-7)	6.05 (-7)	2.27 (-9)
14	3.82 (-9)	6.98 (-9)	1.99 (-11)
16	3.85 (-11)	6.37 (-11)	3.06 (-10)

Spectral Multigrid Methods. Iterative schemes for spectral collocation equations, such as AF1, can be accelerated dramatically by applying multigrid concepts. This technique has been extensively developed for finite difference and finite element discretizations [40] and has recently been applied to spectral discretizations [35,36,29]. Briefly put, multigrid methods take advantage of a property shared by a wide variety of relaxation schemes - potential efficient reduction of the high-frequency error components but unavoidable slow reduction of the low-frequency components.

The fundamentals of spectral multigrid are perhaps easiest to grasp for the simple model problem

$$- \frac{d^2u}{dx^2} = f \tag{76}$$

on $[0,2\pi]$ with periodic boundary conditions. The Fourier approximation to the left-hand side of Eq. (76) at the collocation points is

$$\sum_{p= -N/2+1}^{N/2-1} p^2 \hat{u}_p e^{ipx_j} . \tag{77}$$

The spectral approximation to Eq. (76) may be expressed as

$$LU = F, \tag{78}$$

where

$$U = (u_0,u_1,\cdots,u_{N-1}), \tag{79}$$

$$F = (f_0,f_1,\cdots,f_{N-1}), \tag{80}$$

and L represents the Fourier spectral approximation to $- d^2/dx^2$.

A Richardson's iterative scheme for solving Eq. (78) is

$$V \leftarrow V + \omega(F - LV), \tag{81}$$

where ω is a relaxation parameter. On the right side of the replacement symbol ($\leftarrow$) V represents the current approximation to U, and on the left it represents the updated approximation. The eigenfunctions of L are

$$\xi_j(p) = e^{2\pi ijp/N}, \tag{82}$$

with the corresponding eigenvalues

$$\lambda(p) = p^2, \tag{83}$$

where $j = 0,1,\cdots,N-1$ and $p = -N/2+1,\cdots,N/2-1$. The index p has a natural interpretation as the frequency of the eigenfunction.

The error at any stage of the iterative process is $V - U$; it can be resolved into an expansion in the eigenvectors of L. Each iteration reduces the p'th error component to $\nu(\lambda_p)$ times its previous value, where

$$\nu(\lambda) = 1 - \omega\lambda. \tag{84}$$

The optimal choice of ω results from minimizing $|\nu(\lambda)|$ for $\lambda \varepsilon [\lambda_{min},\lambda_{max}]$, where $\lambda_{min} = 1$ and $\lambda_{max} = N^2/4$. (One need not worry about the $p = 0$ eigenfunction since it corresponds to the mean level of the solution, which is at one's disposal for this problem.) The optimal relaxation parameter for this single-grid procedure is

$$\omega_{SG} = \frac{2}{\lambda_{max} + \lambda_{min}} . \tag{85}$$

It produces the spectral radius

$$\rho_{SG} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} . \tag{86}$$

Unfortunately, $\rho_{SG} \simeq 1 - 8/N^2$, which implies that $O(N^2)$ iterations are required to achieve convergence.

This slow convergence is the outcome of balancing the damping of the lowest-frequency eigenfunction with that of the highest-frequency one in the minimax problem described after Eq. (84). The multigrid approach takes advantage of the fact that the low-frequency modes ($|p| < N/4$) can be represented just as well on coarser grids. It settles for balancing the middle-frequency eigenfunction ($|p| = N/4$) with the highest-frequency one ($|p| = N/2$), and hence damps effectively only those modes which cannot be resolved on coarser grids. In Eqs. (85) and (86), λ_{min} is replaced with $\lambda_{mid} = \lambda(N/4)$. The optimal relaxation parameter in this context is

$$\omega_{MG} = \frac{2}{\lambda_{max} + \lambda_{mid}} . \tag{87}$$

The multigrid smoothing factor

$$\mu_{MG} = \frac{\lambda_{max} - \lambda_{mid}}{\lambda_{max} + \lambda_{mid}} \tag{88}$$

measures the damping rate of the high-frequency modes. In this example $\mu_{MG} = 0.60$, independent of N. The price of this effective damping of the high-frequency errors is that the low-frequency errors are hardly damped at all. Table X compares the single-grid and multigrid damping factors for $N = 64$. However, on a grid with $N/2$ collocation points, the modes for $|p| \varepsilon [N/8, N/4]$ are now the high-frequency ones. They get damped on this grid. Still coarser grids can be used until relaxations are so cheap that one can afford to damp all the remaining modes, or even to solve the discrete equations exactly. For the case illustrated in Table X the high-frequency error reduction in the multigrid context is roughly 250 times as fast as the single-grid reduction for $N = 64$.

Let us consider just the interplay between two grids. A general, nonlinear fine-grid problem can be written

$$L^f(U^f) = F^f. \tag{89}$$

The shift to the coarse grid occurs after the fine-grid approximation V^f has been sufficiently smoothed by the relaxation process, i.e., after the high-frequency content of the error $V^f - U^f$ has been sufficiently reduced. The related coarse-grid problem is

$$L^c(U^c) = F^c, \tag{90}$$

where

$$F^c = R[F^f - L^f(V^f)] + L^c(RV^f). \tag{91}$$

The restriction operator R interpolates a function from the fine grid to the coarse grid. The coarse-grid operator and solution are denoted by L^c and U^c, respectively. After an adequate approximation V^c to the coarse-grid problem has been obtained, the fine-grid approximation is corrected via

$$V^f \leftarrow V^f + P(V^c - RV^f). \tag{92}$$

The prolongation operator P interpolates a function from the coarse grid to the fine grid.

A complete multigrid algorithm requires specific choices of the interpolation operators, the coarse-grid operators, and the relaxation schemes. These issues are discussed at length in [35,36,29] for both Fourier and Chebyshev multigrid methods. Numerous linear, variable coefficient examples are also provided there. The more interesting nonlinear examples from [29] are the subject of the remainder of this paper.

Table X. Damping Factors for N = 64

p	Single-Grid	Multigrid
1	.9980	.9984
2	.9922	.9938
4	.9688	.9750
8	.8751	.9000
12	.7190	.7750
16	.5005	.6000
20	.2195	.3750
24	.1239	.1000
28	.5298	.2250
32	.9980	.6000

Application to Two-Dimensional Potential Flow. Until the recent work of Streett [39], the discretization procedures for the potential equation were invariably based on low-order finite difference or finite element methods. Streett used a spectral discretization of the full potential equation and obtained its solution by a single-grid iterative technique. The application of spectral multigrid techniques by Streett, et al. [29] produced a dramatic acceleration of the iterative scheme. Even in its relatively primitive state the spectral multigrid scheme is competitive, and in some cases unequivocally more efficient, than standard finite difference schemes.

After a conformal mapping from the surface of an airfoil to a circle the potential equation becomes

$$\frac{\partial}{\partial R}\left(R\rho\,\frac{\partial G}{\partial R}\right) + \frac{\partial}{\partial \Theta}\left(\frac{\rho}{R}\,\frac{\partial G}{\partial \Theta}\right) = 0, \tag{93}$$

where G is the reduced potential, R and Θ are the computational polar coordinates, and ρ is the fluid density. The reduced potential is periodic in Θ and it satisfies

$$\frac{\partial G}{\partial R} = 0 \qquad \text{at} \quad R = 1, \tag{94}$$

$$G \to 0 \qquad \text{as} \quad R \to \infty, \tag{95}$$

and the Kutta condition. The density is given by the isentropic relation

$$\rho = \left[1 - \frac{\gamma-1}{2} M_\infty^2(q_r^2 + q_\theta^2 - 1)\right]^{\frac{1}{\gamma-1}}; \qquad (96)$$

the ratio of specific heats is denoted by γ, and M_∞ is the Mach number at infinity. The velocity components in the physical (r,θ) plane are

$$q_r = \frac{1}{H}\frac{\partial\Phi}{\partial R}$$
$$q_\theta = \frac{1}{RH}\frac{\partial\Phi}{\partial\Theta}, \qquad (97)$$

and the Jacobian between the complex physical plane $(z = re^{i\theta})$ and the complex computational plane $(\sigma = Re^{i\Theta})$ is

$$H = \left|\frac{dz}{d\sigma}\right|. \qquad (98)$$

Further details are provided in [39].

The spectral method employs a Fourier series representation in Θ. Constant grid spacing in Θ corresponds to a convenient dense spacing in the physical plane at the leading and trailing edges. The domain in R (with a large, but finite outer cutoff) is mapped onto the standard Chebyshev domain [-1,1] by an analytical stretching transformation that clusters the collocation points near the airfoil surface. The stretching is so severe that the ratio of the largest-to-smallest radial intervals is typically greater than 1000.

The flow past an NACA 0012 airfoil at 4^o angle of attack and a freestream Mach number of 0.5 is a challenging subsonic and thus elliptic case. Nevertheless, the spectral solution on a relatively coarse grid captures all the essential details of the flow. The surface pressure coefficient from the spectral code MGAFSP [29] using 16 points in the radial (R) direction, and 32 points in the azimuthal (Θ) direction is displayed in Fig. 5. The symbols denote the solution at the collocation points. For comparison, the result from the finite difference, multigrid, approximate factorization code FLO36 [41] is shown as a solid line. The grid used in the benchmark finite difference calculation is so fine (64 × 384 points) that the truncation error is well below plotting accuracy. The FLO36 and MGAFSP results are identical to plotting accuracy. The spectral computation on this mesh yields a lift coefficient with truncation error less than 10^{-4}. Spectral solutions on a 16 × 32 grid are thus of more than adequate resolution and accuracy for subsonic flows.

In Fig. 6 are shown convergence histories from FLO36, MGAFSP, and the finite difference, approximate factorization, single-grid code TAIR [42]. Meshes which yield approximately equivalent accuracy were chosen. The surface pressure results are the same to plotting accuracy, the lift coefficient is converged in the third decimal place, and the predicted drag coefficient is less than .001. (Actually, the spectral result is an order of magnitude more accurate than these limits, but the TAIR result barely meets them.) Figure 7 demonstrates the improvement produced by the spectral multigrid scheme over the spectral single-grid method (AFSP). There is well over an order-of-magnitude gain in efficiency.

V. A MIXED EQUATION. The potential flow problem is much more difficult whenever the flow field contains both supersonic (hyperbolic) and subsonic (elliptic) regions. Nevertheless, the spectral multigrid algorithm that succeeded for the subsonic flow case requires only a minor modification in order to succeed for the transonic (mixed) problem as well.

The most expedient technique for dealing with the mixed elliptic-hyperbolic nature of the transonic problem is to use the artificial density approach of Hafez, et al. [43]. The original artificial density is

$$\tilde{\rho} = \rho - \mu \overleftarrow{\delta} \rho \tag{99}$$

with

$$\mu = \max\left\{0, 1 - \frac{1}{M^2}\right\}, \tag{100}$$

where M is the local Mach number and $\overleftarrow{\delta}\rho$ is an upwind first-order (undivided) difference. The spectral calculations employed a higher-order artificial density formula. The spectral method also required a weak filtering technique to deal with some high-frequency oscillations generated by the shock. Details are available in [39].

Flow Past an Airfoil. A lifting transonic case is provided by the NACA 0012 airfoil at $M_\infty = 0.75$ and 2° angle of attack. A shock appears only on the upper surface for these conditions and is rather strong for a potential calculation; the normal Mach number ahead of the shock is about 1.36. Lifting transonic cases are especially difficult for spectral methods since the solution will always have significant content in the entire frequency spectrum: the shock populates the highest frequencies of the grid and the lift is predominantly on the scale of the entire domain. An iterative scheme therefore must be able to damp error components across the spectrum.

Surface pressure distributions from MGAFSP, TAIR, and FLO36 are shown in Fig. 8. The respective computational grids are 18 × 64, 30 × 149, and 32 × 192. The latter two are the default grids for the production finite difference codes. Spectral results obtained by trigonometrically interpolating the 18 × 64 grid results onto a much finer grid are included alongside the results at the collocation points. This reveals the wealth of detail that is provided by the rather coarse spectral grid. The shock predicted by TAIR is far more rounded and smeared than that of FLO36, reflecting the coarser mesh and larger artificial viscosity used in the former. The TAIR result shown is also only correct to one decimal place in lift as compared with a finer-grid result. Convergence histories for these three cases are shown in Fig. 9 along with the results for MGAFSP on a coarser grid (16 × 48).

Flow Past a Circular Cylinder. The MGAFSP code has recently been used for an extremely accurate determination of the critical freestream Mach number at which the potential flow past a circular cylinder first develops a supersonic region [44]. This spectral calculation represents an alternative to the asymptotic series method employed by van Dyke and Guttmann[45] to arrive at the estimate $M_{crit} = .39823780 \pm .00000001$.

The spectral multigrid potential code was used to determine the critical Mach number on several grids. On each of these grids calculations were performed at a half-dozen or so freestream Mach numbers. For each case the maximum local Mach number was determined from the computed solution. Then an extrapolation procedure was employed to ascertain what freestream Mach number produced a maximum local Mach number of unity. This value was recorded as the critical Mach number for that particular grid. An estimate of the extrapolation error was made to ensure consistency. These results are given in Table XI.

Finally, these grid-dependent calculations of the critical freestream Mach number were extrapolated to the limit of infinite numerical resolution. The best result was obtained by assuming sixth-order convergence. The final estimate of the critical freestream Mach number is $M_{crit} = .3982415 \pm .0000002$. The difference between this estimate and the one by van Dyke and Guttmann is more than an order-of-magnitude greater than the estimated errors. Possible explanations for this discrepancy are discussed in [44]. Nevertheless, the agreement of the two estimates to better than one part in 10^5 is remarkable in itself.

Note that the convergence rate of the spectral multigrid potential result (at least sixth-order) pertains to a quantity (critical freestream Mach number) which requires the fundamental solution (the potential) to be first differentiated and then extrapolated. Moreover, the MGAFSP code is so efficient that all of the requisite calculations consumed less than 20 minutes of CPU time on the CDC Cyber 175 and were performed on grids with no more than 2000 points.

A comparable calculation by existing finite difference codes would likely exhibit only first-order convergence. It would be far more expensive both in terms of CPU time and storage, surely exceeding the central memory of a machine such as the CDC Cyber 175. Here then is an example which firmly establishes the utility of spectral methods for nonlinear, multi-dimensional problems.

Table XI. Grid-dependent Critical Freestream Mach Numbers

Grid	M_{crit}	Error Estimate
14 × 32	.398289	.000048
18 × 40	.3982514	.0000099
22 × 48	.3982450	.0000035
30 × 64	.3982422	.0000007

Research of the first author was supported by NASA Contract Nos. NAS1-17130 and NASA-17070 while in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665.

References

[1] Finlayson, B. A. and Scriven, L. E., "The Method of Weighted Residuals - A Review," Appl. Mech. Rev., Vol. 19, 1966, pp. 735-748.

[2] Slater, J. C., "Electronic Energy Bands in Metal," Phys. Rev., Vol. 45, 1934, pp. 794-801.

[3] Barta, J., "Über die Naherungsweise Lösung einiger Zwidimensionaler Elastizitätsaufgaben ," Z. Angew. Math. Mech. (ZAMP), Vol. 17, 1937, pp. 184-185.

[4] Frazer, R. A, Jones, W. P., and Skan, S. W., "Approximation to Functions and to the Solutions of Differential Equations," Great Britain Aero. Res. Council, London, Report and Memo No. 1799, 1937.

[5] Lanczos, C. L., "Trigonometric Interpolation of Empirical and Analytic Functions," J. Math. Phys., Vol. 17, 1938, pp. 123-199.

[6] Clenshaw, C. W., "The Numerical Solution of Linear Differential Equations in Chebyshev Series," Proc. Cambridge Phil. Soc., Vol. 53, 1957, pp. 134-149.

[7] Clenshaw, C. W. and Norton, H. J., "The Solution of Nonlinear Ordinary Differential Equations in Chebyshev Series," Comp. J., Vol. 6, 1963, pp. 88-92.

[8] Wright, K., "Chebyshev Collocation Methods for Ordinary Differential Equations," Comp. J., Vol. 6, 1964, pp. 358-365.

[9] Villadsen, J. V. and Stewart, W. E., "Solution of Boundary Value Problems by Orthogonal Collocation," Chem. Eng. Sci., Vol. 22, 1967, pp. 1483-1501.

[10] Kreiss, H.-O. and Oliger, J., "Comparison of Accurate Methods for the Integration of Hyperbolic Equations," Report No. 36, Department of Computer Science, Uppsala University, Sweden, 1971.

[11] Orszag, S. A., "Comparison of Pseudospectral and Spectral Approximations," Stud. Appl. Math., Vol. 51, 1972, pp. 253-259.

[12] Silberman, I., "Planetary Waves in the Atmosphere," J. Meteor., Vol. 11, 1954, pp. 27-34.

[13] Orszag, S. A., "Numerical Methods for the Simulation of Turbulence," Phys. Fluids, Supplement II, Vol. 12, 1969, pp. 250-257.

[14] Eliasen, E., Machenauer, B., and Rasmussen, E., "On a Numerical Method for Integration of the Hydrodynamical Equations with a Spectral Representation of the Horizontal Fields," Report No. 2, Department of Meteorology, Copenhagen University, Denmark, 1970.

[15] Gottlieb, D. and Orszag, S. A., Numerical Analysis of Spectral Methods: Theory and Applications, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1977.

[16] Salas, M. D., Zang, T. A., and Hussaini, M. Y., "Shock-fitted Euler Solutions to Shock-Vortex Interactions," Proc. of the 8th Intl. Conf. on Numerical Methods in Fluid Dynamics, E. Krause, ed., Springer-Verlag, 1982.

[17] Hussaini, M. Y., Kopriva, D. A., Salas, M. D., and Zang, T. A., "Spectral Methods for the Euler Equations," Proceedings of the Sixth AIAA Computational Fluid Dynamics Conference, Danvers, MA, July 1983.

[18] Gottlieb, D., Lustman, L., and Orszag, S. A., "Spectral Calculations of One-Dimensional Inviscid Compressible Flows," SIAM J. Sci. Statis. Comput., Vol. 2, 1981, pp. 296-310.

[19] Zang, T. A., Hussaini, M. Y., and Bushnell, D. M., "Numerical Computations of Turbulence Amplification in Shock Wave Interactions," AIAA J., to appear.

[20] Hussaini, M. Y., Salas, M. D., and Zang, T. A., "Spectral Methods for Inviscid, Compressible Flows," Advances in Computational Transonics, W. G. Habashi, ed., Pineridge Press, Swansea, UK, 1983.

[21] Pao, S. P. and Salas, M. D., "A Numerical Study of Two-Dimensional Shock Vortex Interaction," AIAA Paper 81-1205, 1981.

[22] Ribner, H. S., "Shock-turbulence Interaction and the Generation of Noise," NACA Report 1233, 1955.

[23] Zang, T. A, Kopriva, D. A. and Hussaini, M. Y., "Pseudospectral Calculation of Shock Turbulence Interactions," Proc. of the 3rd Intl. Conf. on Numerical Methods in Laminar and Turbulent Flow, C. Taylor, ed., Pineridge Press, 1983.

[24] Orszag, S. A. and Patterson, G. S., "Numerical Simulation of Three-Dimensional Homogeneous Isotropic Turbulence," Phys. Rev. Lett., Vol. 28, 1972, pp. 76-79.

[25] Orszag, S. A. and Kells, L. C., "Transition to Turbulence in Plane Poiseuille and Plane Couette Flow," J. Fluid Mech., Vol. 96, 1980, pp. 159-205.

[26] P. Moin and J Kim, "On the Numerical Solution of Time-dependent Viscous Incompressible Fluid Flows Involving Solid Boundaries," J. Comput. Phys., Vol. 35, 1980, pp. 381-392.

[27] Kleiser, L. and Schumann, U., "Spectral Simulation of the laminar-turbulent transition process in plane Poiseuille flow," Proc. of ICASE Symp. on Spectral Methods, R. Voigt, ed., SIAM-CBMS, 1983.

[28] Davis, P. S. and Rabinowitz, P., Numerical Integration, Blaisdell Publishing Company, 1967.

[29] Streett, C. L., Zang, T. A., and Hussaini, M. Y., "Spectral Multigrid Methods with Applications to Transonic Potential Flow," ICASE Report No. 83-11, 1983.

[30] Lanczos, C. "Legendre versus Chebyshev Polynomials," Proc. of the Royal Irish Academy Conference on Numerical Analysis, John J. H. Milles, ed., Academic Press, 1973.

[31] Haidvogel, D. B. and Zang, T. A., "The Accurate Solution of Poisson's Equation by Expansion in Chebyshev Polynomials," J. Comput. Phys., Vol. 30, 1979, pp. 167-180.

[32] Malik, M. R., Zang, T. A. and Hussaini, M. Y., "Efficient Solution to Semi-implicit Spectral Methods for Navier-Stokes Equations," to appear.

[33] Morchoisne, Y., "Resolution of Navier-Stokes Equations by a Space-time Pseudospectral Method," La Recherche Aerospatial, Vol. 5, 1979, pp. 293-309.

[34] Orszag, S. A., "Spectral Methods for Problems in Complex Geometries," J. Comput. Phys., Vol. 37, 1980, pp. 70-92.

[35] Zang, T. A., Wong, Y. S., and Hussaini, M. Y., "Spectral Multigrid Methods for Elliptic Equations," J. Comput. Phys., Vol. 48, 1982, pp. 485-501.

[36] Zang, T. A., Wong Y. S., and Hussaini, M. Y., "Spectral Multigrid Methods for Elliptic Equations II," NASA CR 172131, 1983

[37] Douglas, J. and Gunn, J. E., "A General Formulation of Alternating Direction Method," Numer. Math., Vol. 6, 1964, pp. 428-453.

[38] Ballhaus, W. F., Jameson, A., and Albert, J., "Implicit Approximate Factorization Schemes for the Efficient Solution of Steady Transonic Flow Problems," AIAA J., Vol. 16, 1978, pp. 573-579.

[39] Streett, C. L., "A Spectral Method for the Solution of Transonic Potential Flow About an Arbitrary Airfoil," Proc. of the Sixth AIAA Computational Fluid Dynamics Conference, Danvers, MA, July 1983.

[40] Hackbusch, W. and Trottenberg, U., eds., Multigrid Methods, Lecture Notes in Mathematics 960, Springer-Verlag, New York, 1982.

[41] Jameson, A., "Acceleration of Transonic Potential Flow Calculations on Arbitrary Meshes by the Multiple Grid Method, " AIAA Paper 79-1458, 1979.

[42] Holst, T. L., "A Fast, Conservative Algorithm for Solving the Transonic Full-Potential Equation," AIAA Paper 79-1456, 1979.

[43] Hafez, M. M., South, J. C., and Murman, E. M., "Artificial Compressibility Methods for Numerical Solution of Transonic Full Potential Equation," AIAA J., Vol. 17, 1979, pp. 838-844.

[44] Streett, C. L., Zang, T. A., and Hussaini, M. Y., to appear.

[45] van Dyke, M. D. and Guttmann, A. J., "Subsonic Potential Flow Past a Circle and the Transonic Controversy," J. Austral. Math. Soc., Ser. B24, 1983, pp. 243-261.

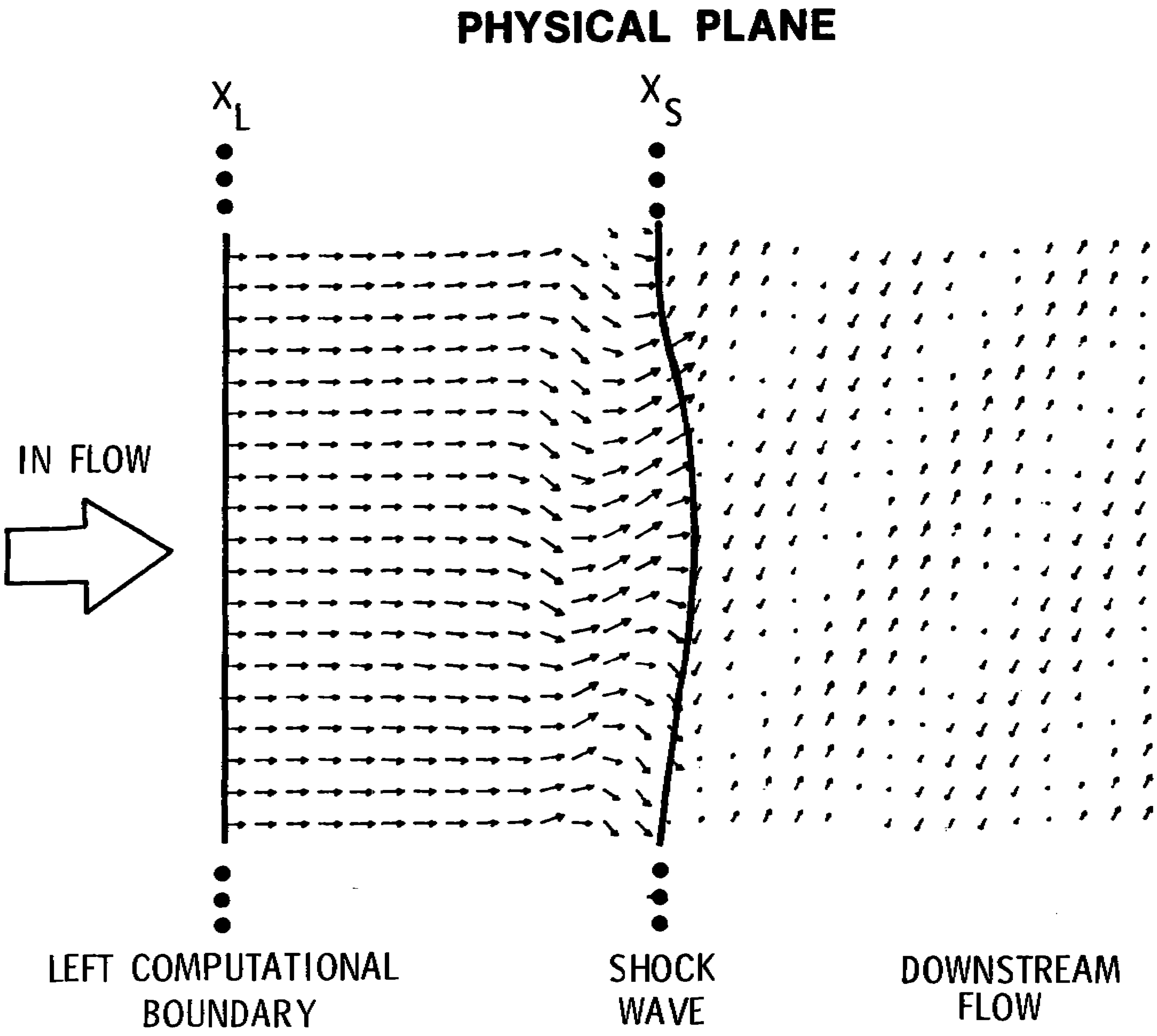


Figure 1
Typical shock-fitted time-depentent flow model in the physical plane.

ACOUSTIC RESPONSE TO 10° INCIDENT ACOUSTIC WAVE

Mach 3

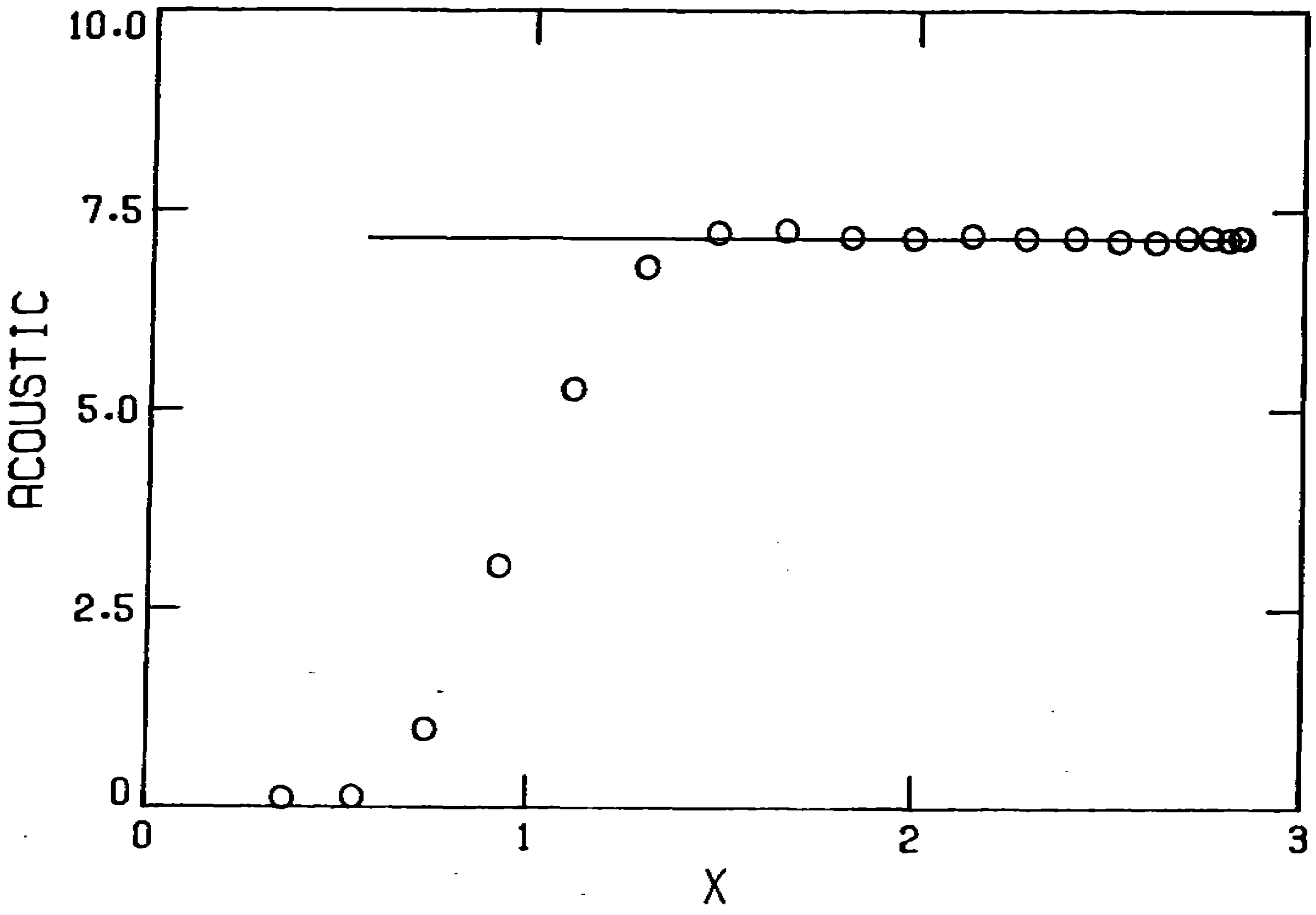
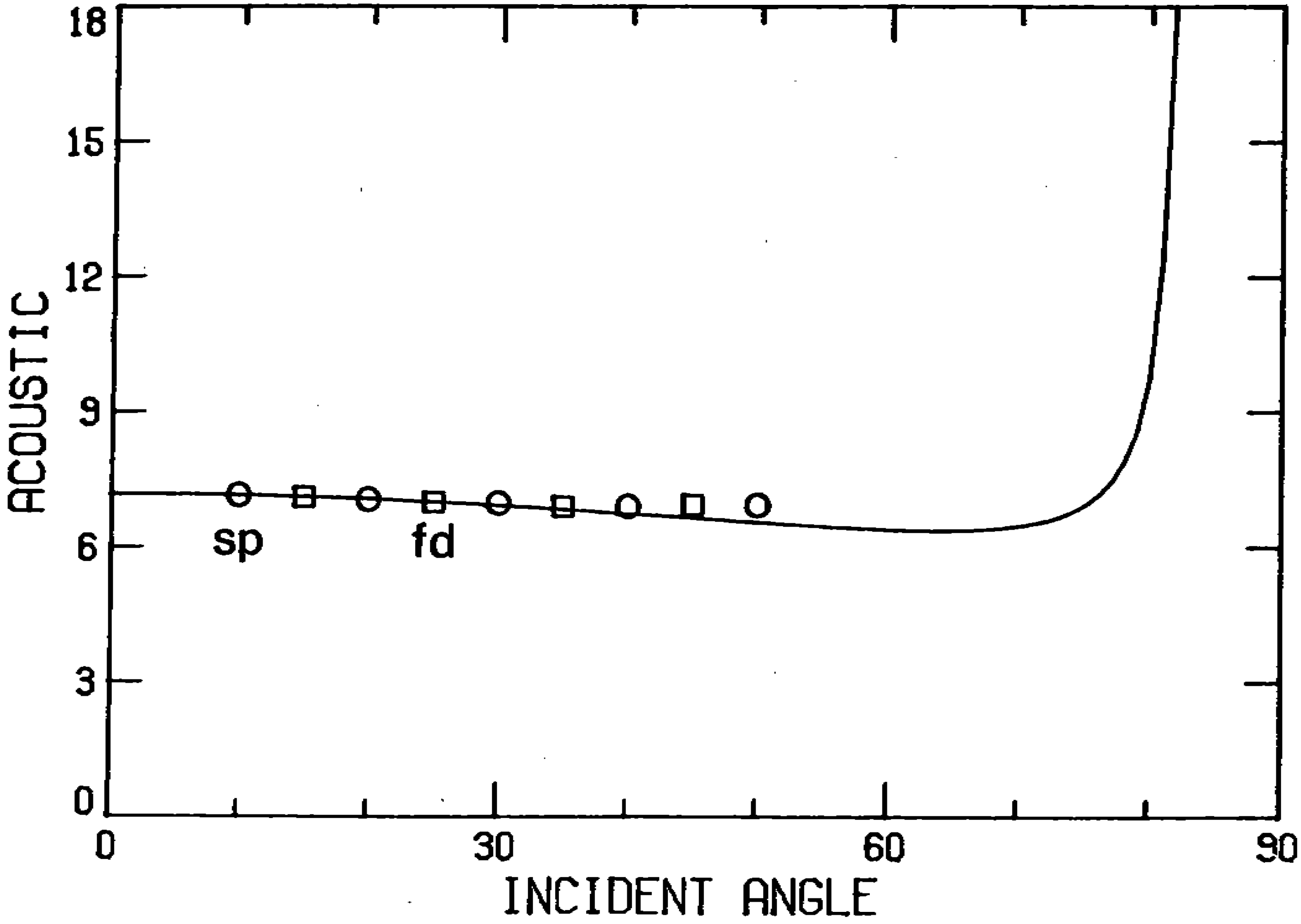


Figure 2

Post-shock dependence of the pressure resonse to a pressure wave incident at 10° to a Mach 3 shock. The solid line is the linear theory prediction. The circles are the spectral solition.

ACOUSTIC TRANSMISSION COEFFICIENT FOR INCIDENT ACOUSTIC WAVES

Mach 3



Dependence on incident angle of the pressure responce to a 0.1% amplitude pressure wave incident on a Mach 3 shock. The solid line is the linear theory result. Circles are spectral solutions; squares are finite difference solutions.

Figure 3

K = 5 CHANNEL FLOW EIGENVALUES

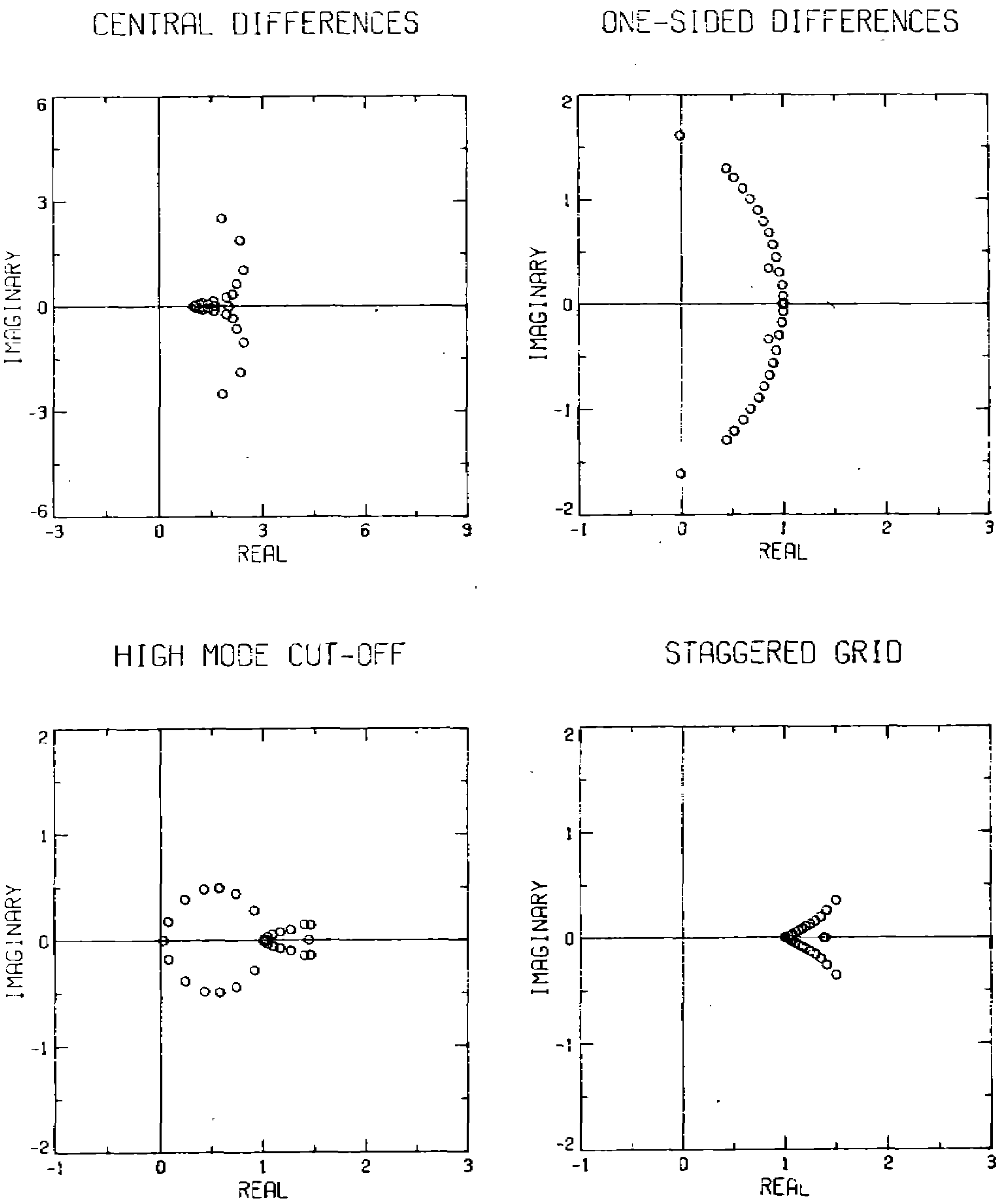


Figure 4

Eigenvalues of the pre-conditioned matrices for the semi-implicite channel flow when the streamwise wave number k = 5. The grid is 32 × 17, the Reynolds number is 7500 and the CFL number is 0.10. Note the different scale used for the central differences pre-conditioning results.

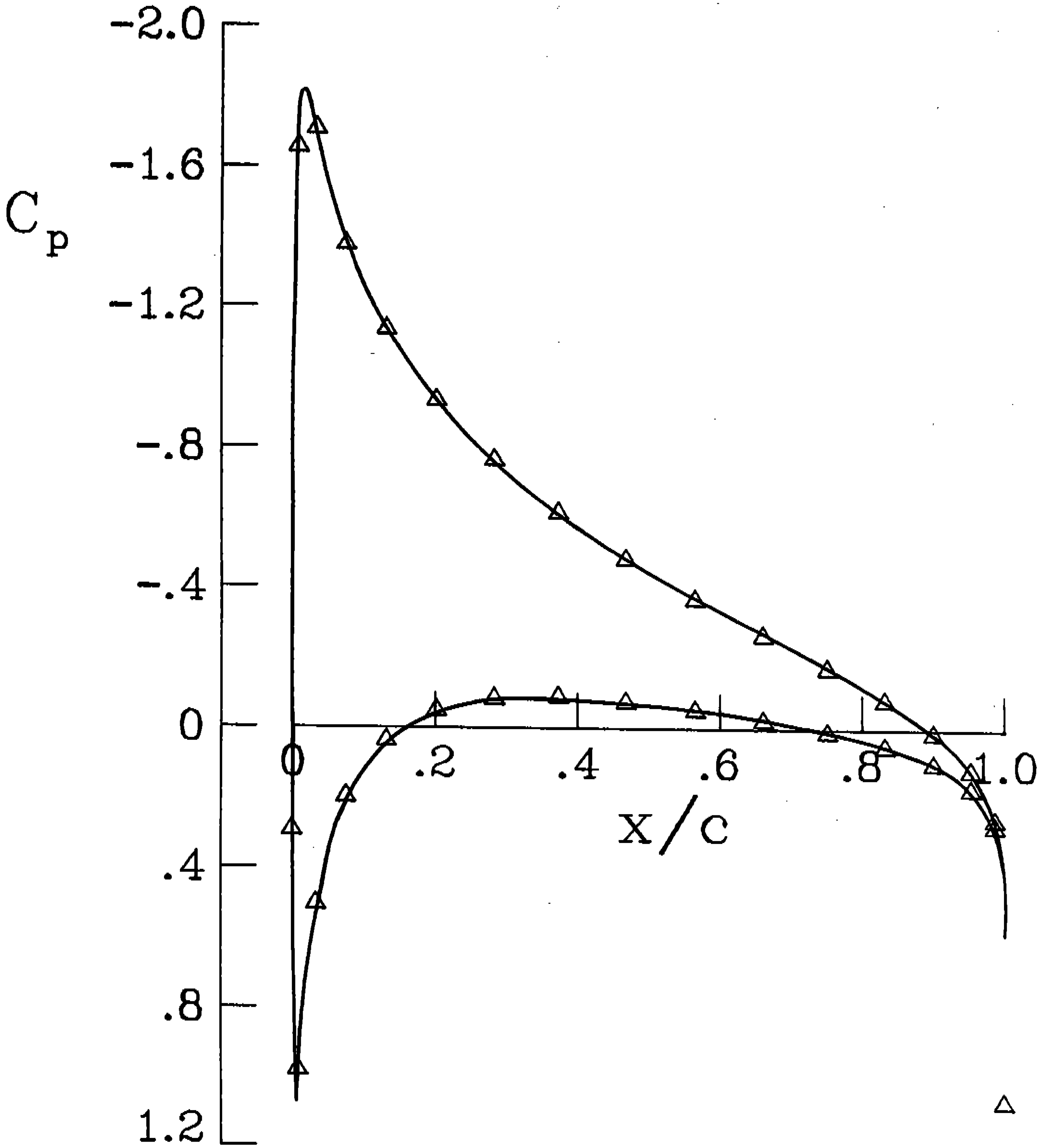


Figure 5
Spectral (triangles) and finite difference (solid line) surface pressures for a subcrital flow.

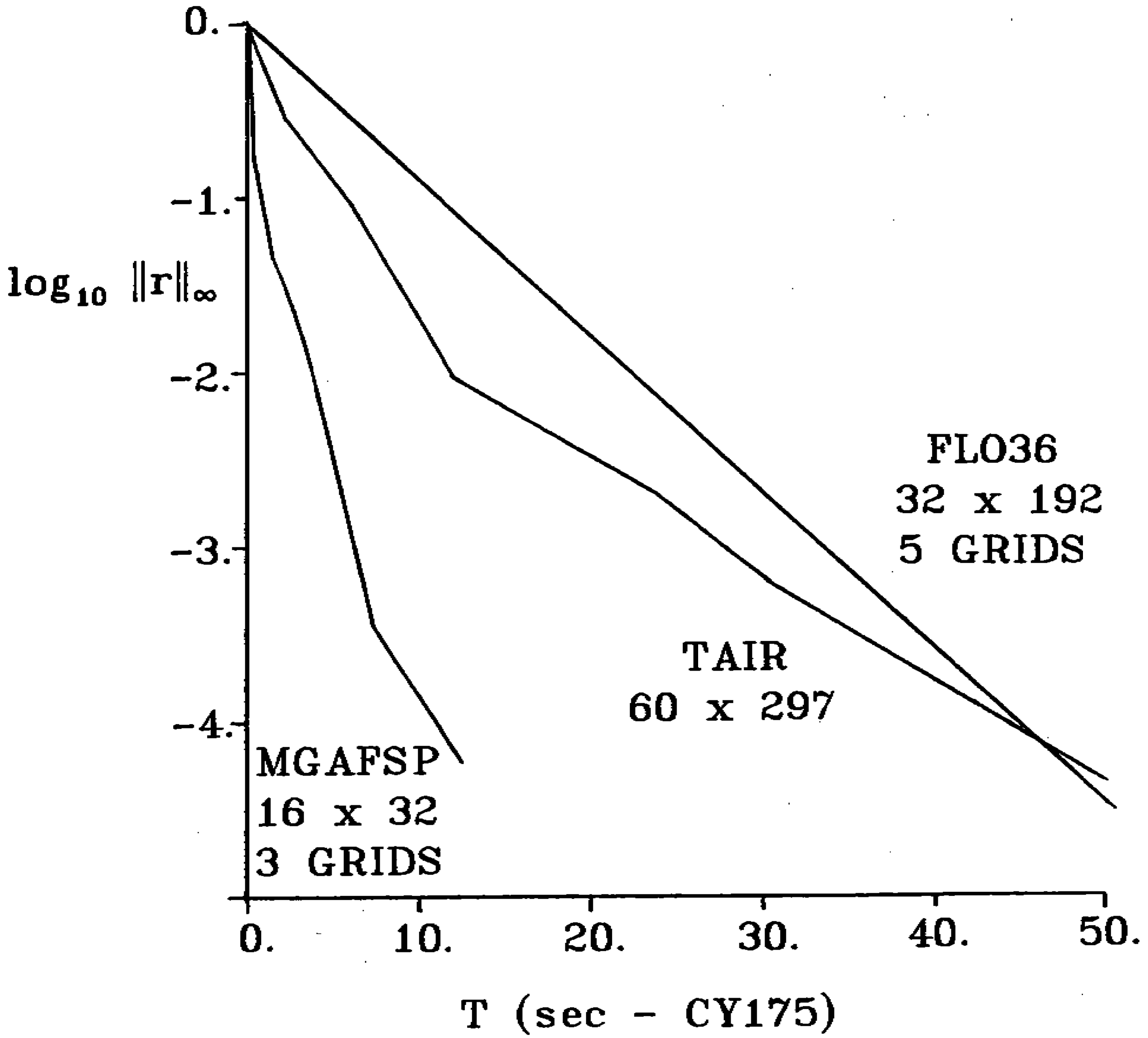


Figure 6

Maximum residual versus machine time foe a subsonic flow

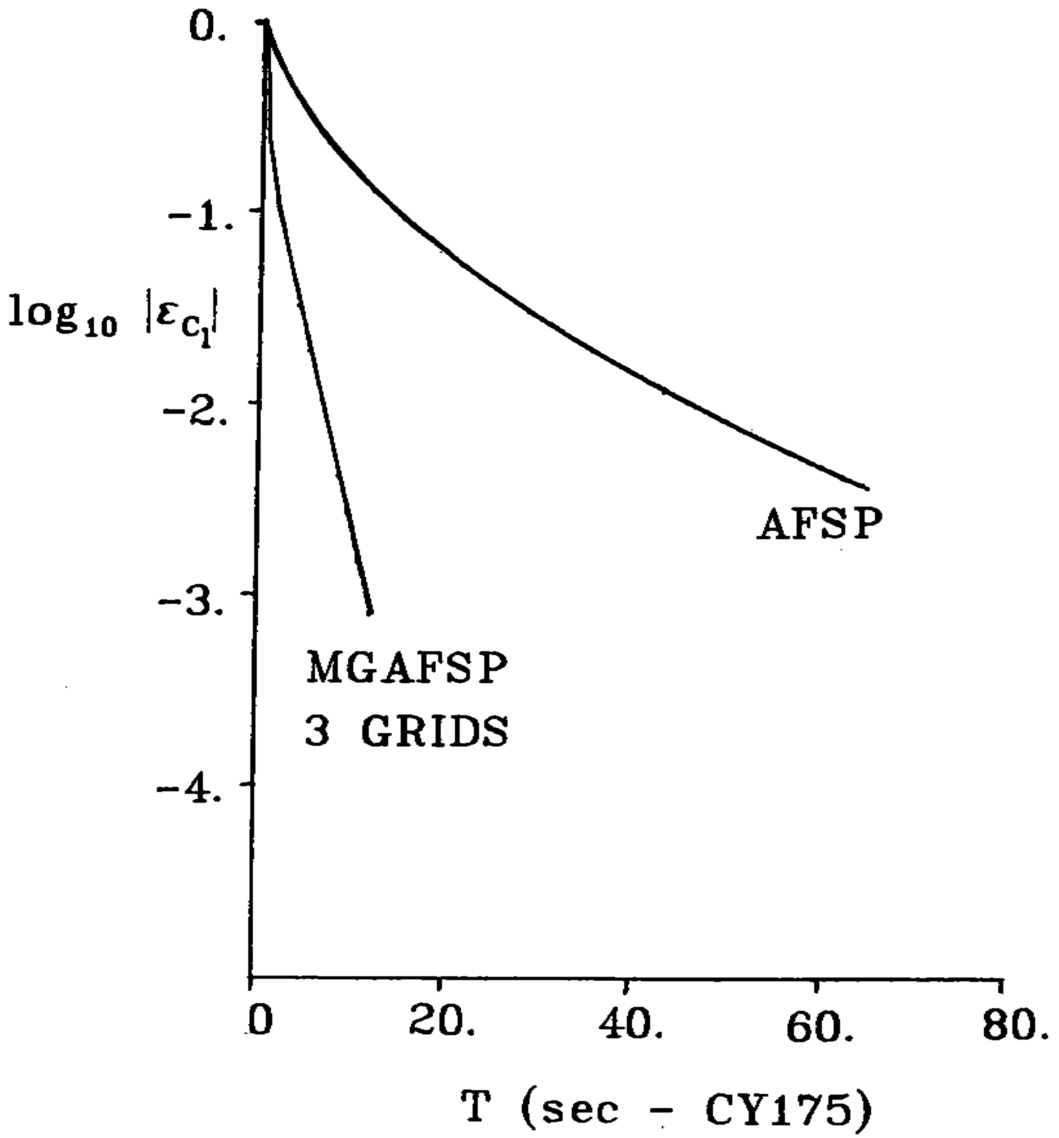


Figure 7

Error in lift versus machine time for a subsonic flow from single-grid (AFSP) and multigrid (MGFSP) spectral schemes.

LIFTING, SUPERCRITICAL POTENTIAL FLOW

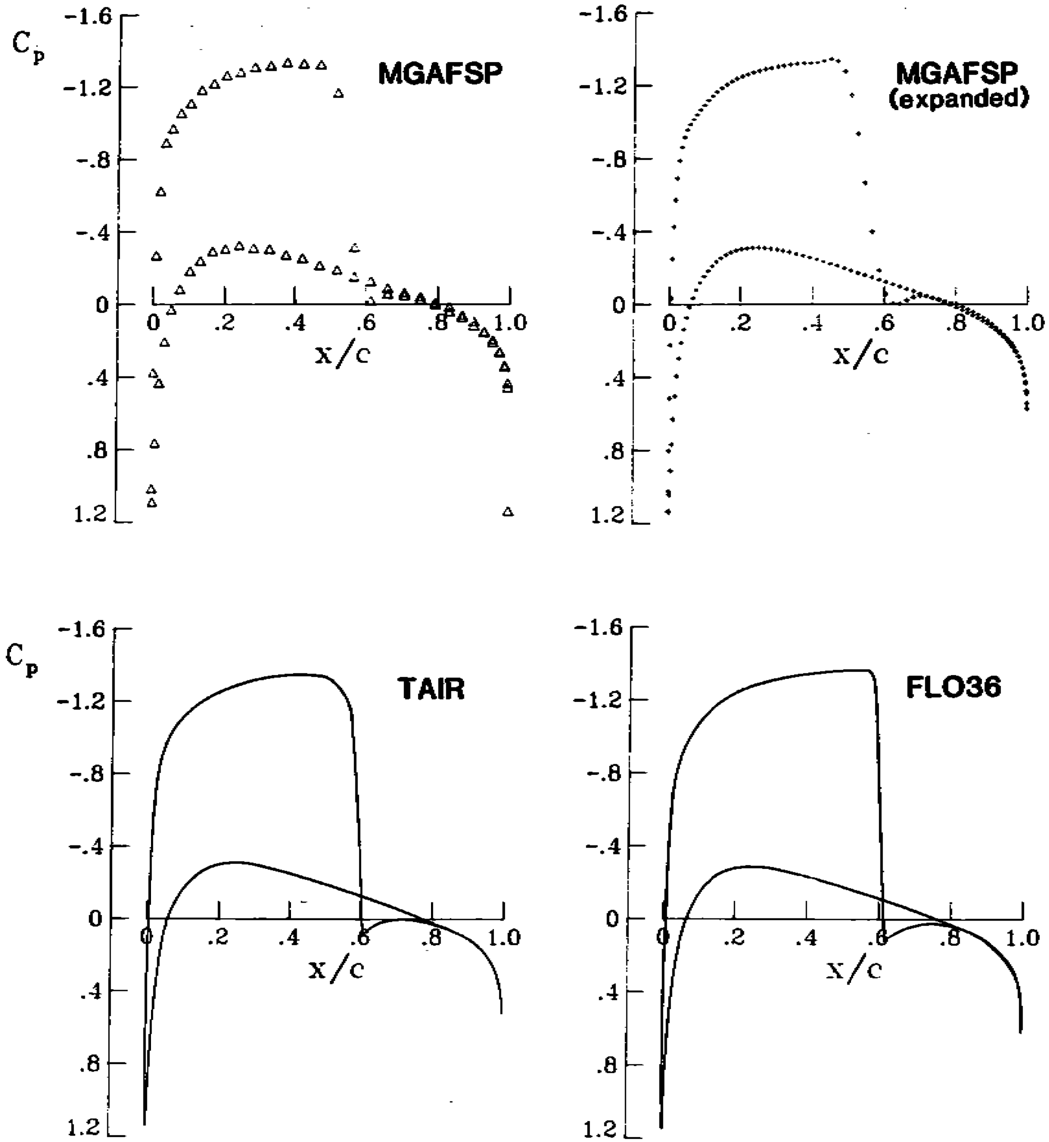


Figure 8
Surface pressures for a transonic flow.

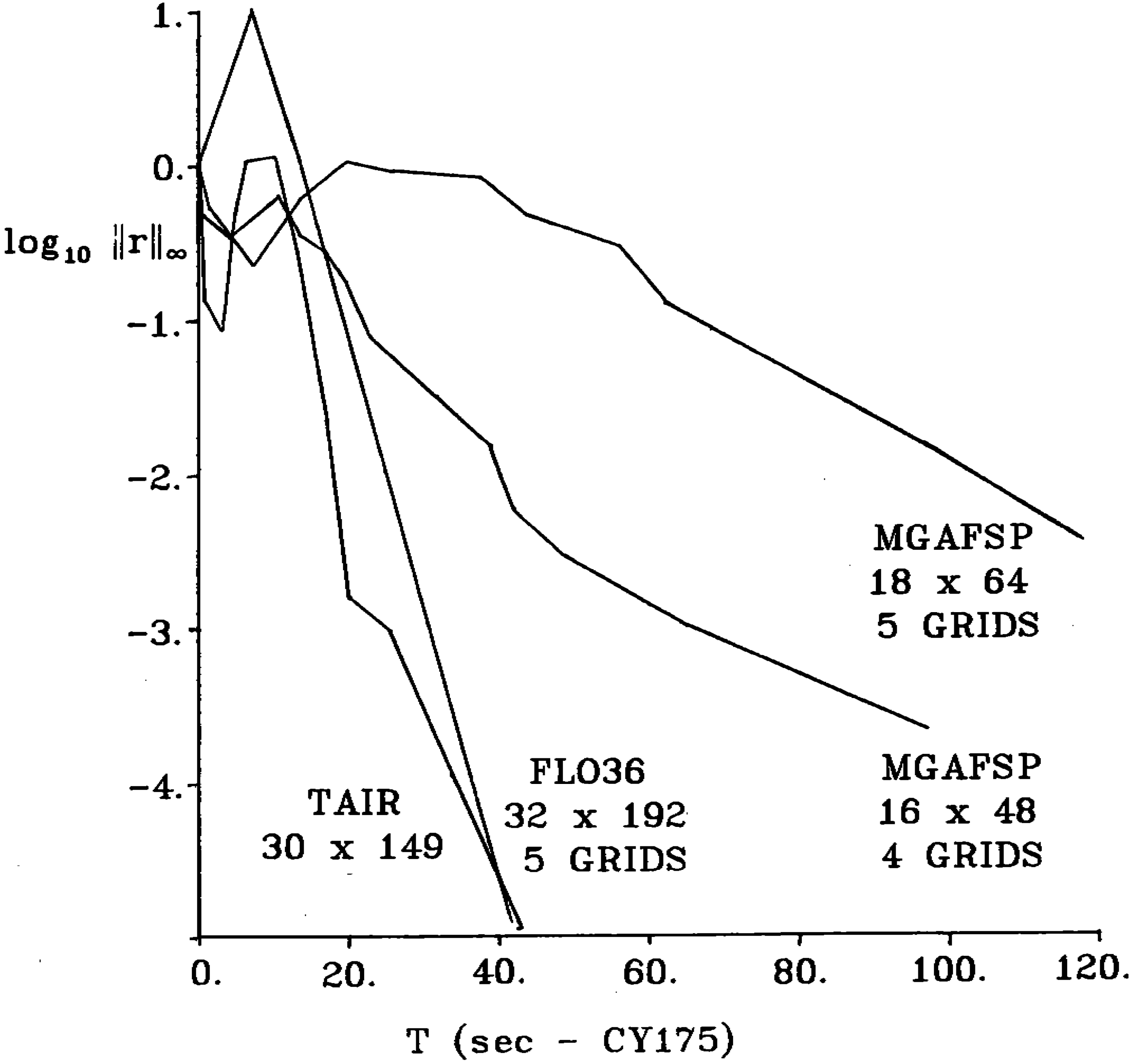


Figure 9
Maximum residual versus machine time for a transonic flow.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
ARO Report 84-1

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Transactions of the First Army Conference on Applied Mathematics and Computing

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Army Mathematics Steering Committee on behalf of the Chief of Research, Development and Acquisition

12. REPORT DATE
February, 1984

13. NUMBER OF PAGES
925

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

STRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. The findings in this report are not to be construed as official Department of the Army position unless so designated by other authorized documents.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
This is a technical report resulting from the First Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat various Army applied mathematical problems.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

asymptotics and numerics
flow problems
mechanical systems
conversion of LINPACK to ADA
Taylor series in PASCAL-SC
transient responses
tracking schemes
Riemann solvers
stochastic systems
group methods
nonlinear integral equations
compustion problems
semiconductors
wave scattering
least squares and residuals
trajectory estimation
discrete-time linear systems
target modeling
structural optimization
stress analysis
finite element and finite difference methods
geometric and dynamic programming
C^3I systems
fuzzy situations
differential equations
Hermite basis functions
splines

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)